SAPIENZA
UNIVERSITÀ DI ROMA

Integrating causal graphs and potential outcomes: Theory, applications and a novel method

Department of Methods and Models for Territory, Economics and Finance
Economic Statistics (XXXIV cycle)

Lorenzo Giammei
ID number 1841755

Advisor
Prof. Paola Vicard

Academic Year 2021/2022

Integrating causal graphs and potential outcomes: Theory, applications and a novel method

PhD thesis. Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: lorenzo.giammei@uniroma1.it

Contents

Premise

Causality is a recurrent theme in economic literature. Evaluating the effect of policies is often the motivational input to estimate causal effects, usually through methods such as matching, difference in differences, regression discontinuity or instrumental variables. All the mentioned techniques belong to the Potential Outcomes framework, one of the main approaches to deal with causality.

The other main causality approach is the Causal Graphs framework. Causal graphs are commonly employed in biomedical sciences, particularly epidemiology, and they contributed to developing a higher awareness of how causal inference is made and how to assess the validity of the findings. Causal graphs had such a substantial impact on some disciplines that the implementation of the relative methods is sometimes referred to as a causal revolution. Nevertheless, causal graphical models did not meet the same consensus in economics, where instead, the implementation of the framework in empirical applications is still sporadic.

Could economics, and other social sciences where potential outcomes methods are standard practice, also benefit from causal graphs? Are those disciplines missing an opportunity by only resorting to potential outcomes? Are the two frameworks conflicting, or can they benefit from a combined implementation? These kinds of questions motivate the research carried out in this thesis.

The thesis consists of three chapters that contain three separate papers. The first paper focuses on investigating if and how potential outcomes and causal graphs are compatible on a theoretical level. The main concepts of the two frameworks are outlined, and their similarities and complementarities highlighted. The second paper takes a step forward, carrying out an empirical economic study in which causal graphs and potential outcomes are employed together. Finally, the last paper proposes a novel methodology that aims at increasing the reliability of causal estimates when subject matter causal knowledge is not available or partial. The procedure is based on causal graphs but is also consistent with potential outcome methods.

Chapter 1

An integrated approach to causality: The role of causal graphs

Abstract

Causal questions are central for most biomedical and social science studies. The main frameworks that allow the analysis of causal relations are Potential Outcomes and Causal Graphs. The approaches have often been compared, contrasting their relative strengths. This paper evaluates the implications of merging the two methodologies in an integrated approach. In particular, we assess how the limits of one can be compensated by the solutions provided by the other. The outlined approach employs causal graphs to discover and formalize a causal model that is then used as a guide to implementing potential outcomes identification strategies. The integrated approach could be beneficial to both frameworks. The assumptions required by potential outcome methods can be assessed directly from a causal graph even in high dimensional contexts, thus making the obtained causal estimates more reliable. On the other hand, causal graphs can benefit from the several ad hoc identification strategies that have been developed in the potential outcomes literature.

1.1 Introduction

The study of cause and effect relations motivates most research in social, demographic and health sciences. Investigating causality usually means assessing if and how a certain intervention, often called treatment, affects an outcome of interest. The early work of Neyman and Iwaszkiewicz (1935), Fisher (1949) and Cox (1958) in the field of randomized experiments constituted a first step towards a rigorous analysis of causality. Based on these studies Rubin (1974) formalizes one of the most relevant approaches to causality: the *Potential Outcomes* (PO) framework. The framework has then been enriched with many contributions that proposed new methods and applications (Imbens and Rubin 2015; P. Rosenbaum 2018). PO have a strong connection with economics since its early stages as its concepts are rooted in the work of Tinbergen (1930) and Haavelmo (1943). PO methods are now widely applied in statistics and economics and many econometric textbooks solely rely on this approach (Angrist and Pischke 2008; Imbens and Rubin 2015).

The other main approach to deal with causality is the *Causal Graph* framework. Note that causal graphs, also called causal bayesian networks or causal diagrams, can be seen as part of a wider model called structural causal model (SCM) (Pearl 2000). In a SCM the causal graph is also associated to a set of equations that describe causal relations between the nodes of the graph. Here we will however only focus on the causal graph component, that is sufficient for answering causal queries concerning the effect of interventions.

Causal graphs have been introduced by Pearl (1995) and share some elements with the previous work on path diagrams in Wright (1921b). The framework have been subsequently developed and enriched with several contributions that extended its applicability and strengthened its results (Pearl 2000; Tian and Pearl 2002; Bareinboim and Pearl 2016; Huang and Valtorta 2012). Causal graphs are now frequently used in epidemiology, computer science and some social sciences, though they are still uncommon in economics.

The relative advantages of the two frameworks have been recently reviewed and compared in Imbens (2020) and Hünermund and Bareinboim (2019). Both papers show some specific causal problems where one approach is more appropriate than the other and vice-versa, thus revealing that, at least in part, the two are complementary and could benefit from each other. The idea of an integrated approach also starts to appear in some causal inference textbooks, such as Morgan and Winship (2015) and Cunningham (2021), however integrated applications are still very rare in practice.

In this paper, we assess how PO and Causal Graphs can be combined and the implications of carrying out such an approach. The basic ideas of the frameworks will be described focusing on when the limits of one way of proceeding are compensated by the other. Particular attention will be put on causal discovery techniques, a resource that is often overlooked when comparing PO and causal graphs. Throughout the paper, we provide some basic examples in which the combination of the frameworks can improve the results' quality and reliability.

Section 1.2 will outline the PO framework, its main assumptions, results and limits. Section 1.3 is instead devoted to Causal Graphs. The basic terminology is presented and the principal features are described with the help of some examples. Then we show how causal effect estimation can be performed from causal graphs and how the process can be integrated with PO methods. Finally, in section 1.4 we introduce the concept of causal discovery; we explain how structural learning algorithms work and why they can be valuable for causal effect estimation.

1.2 Potential Outcomes

The Potential Outcomes (PO) framework originates from the work of Splawa-Neyman, Dabrowska, and Speed (1990) and Rubin (1974) on randomized controlled trials (RCT). The name of the framework comes from its peculiar notation $Y_i(t)$ that denotes the *potential outcome* for unit i when receiving the treatment level $T = t$. In the case of a binary treatment T takes value 1 if unit i is treated and 0 otherwise. Accordingly, $Y_i(1)$ represents the PO we would observe for unit i if it was treated and $Y_i(0)$ the potential outcome if unit i was a control. The causal effect of T on

Y can therefore be computed by comparing summary statistics of the potential outcomes distribution. The resulting causal estimate is usually called the average treatment effect (ATE) and can be expressed in different ways, such as

$$ATE = E[Y_i(1) - Y_i(0)] \qquad \text{or} \qquad ATE = \frac{E[Y_i(1)]}{E[Y_i(0)]}.$$

However, the ATE cannot be estimated directly from data since only one of the potential outcomes is observed for each unit i. Units receive only one level of treatment, creating a missing data problem. This is sometimes referred to as the fundamental problem of causal inference (Holland 1986).

PO literature contributed to answering this problem in the context of randomized experiments. In this setting, treatment is assigned randomly to the units of the sample, thus rendering T independent of the potential outcomes

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)).$$

This scenario, together with the assumption that there is no interference between units (SUTVA)(Imbens and Rubin 2015), ensure that an unbiased estimate of the ATE can be obtained by computing the difference

$$\bar{Y}_t - \bar{Y}_c, \qquad \text{with} \qquad \bar{Y}_t = \frac{1}{N_t}\sum_{i:T_i=1} Y_i \qquad \text{and} \qquad \bar{Y}_c = \frac{1}{N_c}\sum_{i:T_i=0} Y_i.$$

The indexes $i : T_i = t$ indicate to sum over the units that received a certain treatment level, N_t and N_c denote respectively the number of treated and control units.

The PO framework also provides several solutions to deal with non-experimental or observational data. What usually prevents observational data from being treated as experimental data is the presence of *confounders*. Confounders are variables that affect both the treatment and the outcome and can lead to biased causal estimates if not adequately accounted for. The concern worsens when confounders are unobserved since, in this situation, treatment effects could be impossible to identify.

PO methods that deal with observational data aim at emulating an experimental context under specific assumptions. One of these assumptions, that tackles directly the problem of confounders, is called unconfoundedness or ignorability and can be defined as follows

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) | X_i$$

where X_i is a set of pre-treatment covariates. Unconfoundedness states that the treatment T_i is independent of the potential outcomes, given a set of pre-treatment variables X_i. The condition allows estimating the ATE as

$$ATE = E[Y_i(1) - Y_i(0)] = E[E[Y_i|T_i = 1, X_i] - E[Y_i|T_i = 0, X_i]]. \tag{1.1}$$

The formula in Equation 1.1 is also called adjusting for X and as long as unconfoundedness holds, it ensures an unbiased estimation of the ATE in the presence of confounders. Adjustment can be performed through various methods, including regression, matching and inverse probability weighting.

Another PO method to derive causal estimates from observational data is the instrumental variable (IV) strategy (Angrist 1990). In this context, there is an unobserved variable U, which violates the unconfoundedness assumption for the effect of T on Y. Since U is unobserved, it is impossible to adjust for it in order to obtain unbiased estimates. However, if the treatment T is affected by another variable Z, it is still possible to estimate a causal effect, under an assumption called *exclusion restriction.* The assumption can be expressed as

$$Y_i(z, t) = Y_i(z', t) \qquad \text{for all } z, z',$$

imposing that potential outcomes do not vary with Z. PO literature refers to variables that satisfy the exclusion restriction as instrumental variables.

However, exclusion restriction and unconfoundedness cannot be tested, and they are usually motivated by background theory concerning the causal relations between variables. This implies that justifying them becomes difficult if a priori knowledge is missing. Moreover, as the number of variables in the model increases, assessing the two assumptions' validity turns out to be a challenging task.

The PO framework includes many more identification strategies, such as difference-in-differences, regression discontinuity and synthetic control. For a review of the newest techniques, see Athey and Imbens (2017). These methods provide solutions to very specific causal problems and usually impose additional functional-forms restrictions on probability distributions, such as linearity, monotonicity or additivity.

1.3 Causal graphs

In this section, the Causal Graph framework will be described. First, we will introduce the basic terminology of graphs and the main elements of causal graph theory. Next, we will show how interventions are represented in the framework and how causal effects can be estimated employing graphs.

1.3.1 Terminology and basic concepts

A *graph* $G = (V, E)$ is a collection of vertices or nodes V and edges E. The edges can be directed or indirected. An edge that goes from a vertex V_i to another vertex V_j is a *directed edge.* Conversely, an edge without such orientation is an *undirected edge.* A graph that only contains directed edges is called a *directed graph.* When two nodes are connected by an edge, they are called *adjacent* nodes. If each pair of nodes belonging to V is connected by an edge, the graph is called a *complete graph.* Conversely, if none of the pairs is adjacent, the graph is an *empty graph.* A sequence of connected edges that starts from a node V_i and ends with node V_j, regardless of the directions of the edges, is called a *path.* In a *directed path* all the edges are oriented in the same direction along the path. A directed path, starting from V_j and ending in V_i, with $V_j = V_i$ is a *cycle.* A directed graph that contains no cycles is also called a *directed acyclic graph* (DAG)(Pearl 2000). In the context of causal graphs, DAGs are employed to represent causal structures. The vertices of the DAG represent random variables, and its edges describe the causal relations between them. We will refer to variables and vertices in a DAG interchangeably from now on.

Consider the graph G in Figure 1.1. All the edges in the graph are directed, and they form no cycles; the graph is, therefore, a DAG. G describes the multivariate causal relations between a set of four random variables $\mathbf{X}$. The terminology of kinship is often used to indicate relationships between nodes according to the graph's structure.

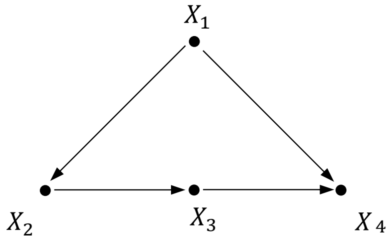

Figure 1.1. A simple DAG (1)

Since the DAG contains a directed edge going from X_1 to X_2, X_1 is called a *parent* of X_2 and the latter is a *child* of X_1. The path p along the ordered sequence of nodes (X_1, X_2, X_3, X_4) is a directed path since all the edges are oriented in the same direction along the path. X_1 is called an *ancestor* of each node belonging to$\{X_2, X_3, X_4\}$ since it precedes them in p and the vertices in $\{X_2, X_3, X_4\}$ are *descendants* of X_1. Given that the edges are carriers of causal information, we can also say that X_1 is a direct cause of X_2 and X_4. The same is true for every ordered pair of random variables (X_i, X_j) connected by a directed edge that goes from X_i

to X_j in the DAG.

Every causal graph also consists of a joint probability distribution $P(\mathbf{X})$ over the variables described by the DAG. This distribution can be factorized according to the structure of the DAG as

$$P(x_1, \dots, x_n) = \prod_i P(x_i | pa_i), \tag{1.2}$$

where pa_i indicate the parent set of variable X_i. The factorization implies that given a DAG G with node set $\mathbf{X}$, for each variable $X_i \in \mathbf{X}$, its parent set PA_i selected according to the structure of G, is sufficient for determining the probability of X_i. If a probability function P admits the factorization of Equation 1.2 relative to a DAG G, then G is said to satisfy the *causal Markov condition* and P is said to be Markov relative to G.

1.3.2 Edge configurations and conditional independence

The edges of a DAG can assume specific configurations that provide additional information regarding the independence relations among variables of the model. Given the ordered triplet of nodes (X_i, X_j, X_k), if two directed edges goes from X_i and X_k to X_j but X_i and X_k are not adjacent, then X_j is called a *collider* or unshielded collider in the ordered triplet. Colliders are also referred to as non-emitting nodes. Conversely, given a path p, vertices belonging to p with at least an outgoing edge directed towards other adjacent nodes in p are called *emitting nodes.* An example of a configuration that only contains emitting nodes is when a directed edge goes from node X_j to node X_i, and another directed edge goes from X_j to a third node X_k. This configuration is called *chain.* Note that the same node can have different roles when considering different paths. Consider the graph in Figure 1.2, in the path along the ordered triplet of nodes (X_3, X_4, X_2), X_4 is a collider, whereas in the path formed by the ordered triplet (X_3, X_4, X_5) the same node is an emitting node.

A DAG encodes information concerning conditional independence among the variables it represents through a criterion called *d-separation.* Consider a DAG G with node set $\mathbf{X}$, a pair of nodes $\{X_i, X_j\}$ belonging to $\mathbf{X}$ with $X_i \neq X_j$ and a set of nodes $S \subset \mathbf{X}$ not containing X_i and X_j. A path p between X_i and X_j is said to be blocked by a set $\mathbf{S}$ in G, if either

1. p contains at least one emitting node that belongs to $\mathbf{S}$
2. p contains at least a collider that does not belong to $\mathbf{S}$ and has no descendent in $\mathbf{S}$.

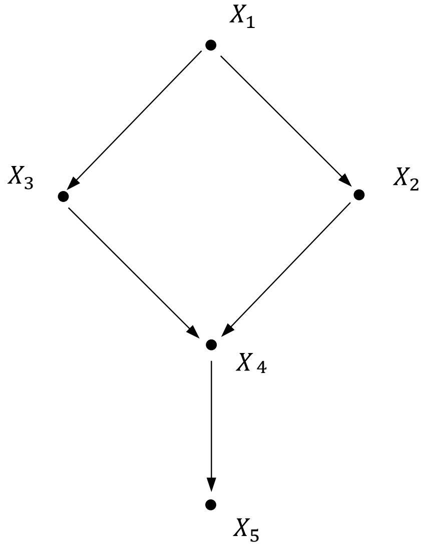

Figure 1.2. A simple DAG (2)

Two nodes X_i and X_j are said to be *d-separated* given a set $\mathbf{S}$ if all the paths between the nodes are blocked by $\mathbf{S}$. When two nodes X_i and X_j are d-separated by a set $\mathbf{S}$, then X_i is independent of X_j conditional on $\mathbf{S}$. Note that two nodes can also be d-separated conditioning on an empty set if all the paths between them contain at least a collider or its descendants. In this case, the variables represented by the nodes are said to be marginally independent. Consider the vertices pair $\{X_2, X_3\}$ in the DAG of Figure 1.2 and the two paths connecting the nodes. The path along the ordered triplet (X_3, X_1, X_2) can be blocked by conditioning on the middle node X_1, since it is an emitting node. The second path, traced along the ordered triplet (X_3, X_4, X_2) is blocked by the collider X_4, without performing any conditioning. The pair $\{X_3, X_2\}$ is thus d-separated by $\mathbf{S} = \{X_1\}$ because conditioning on X_1 blocks every path between the nodes of the pair. The set $\{X_1, X_5\}$ instead, does not d-separate X_2 from X_3 because conditioning on X_5 opens the colliding path along the nodes (X_3, X_4, X_2).

1.3.3 Causal graph analysis at interventional level

Causal graphs allow estimating the effect of interventions, or in other words, the effect of forcing a variable to take a certain value by an external action. Pearl (2000) introduces the *do-operator* $do(X = x)$, a notation to indicate that a variable X is forced by intervention to take value x. In order to be coherent with the terminology defined in Section 1.2 for the PO framework, we will refer to the effect of a treatment variable T on an outcome variable Y. The do-operator allows writing $P(Y|do(T = t))$ to denote the distribution of Y given an intervention that sets $T = t$. This is different form $P(Y|T = t)$ that instead represents the observational distribution of Y given $T = t$. The causal effect of T on Y can thus be obtained by comparing the quantity $P(Y|do(T = t))$ for different values of t, similarly to what is done in the PO framework where instead $Y(t)$ was the quantity of interest. However, when dealing with non-experimental data, causal effects cannot be estimated directly from data since the interventional distribution of Y is not an observed quantity.

Backdoor criterion

One of the critical contributions of causal graphs is that their structure can serve as a guide to express interventional distributions in terms of observational quantities, thus making it possible to estimate causal effects. This is a crucial result since conditional distributions such as $P(Y|T = t)$, can be directly computed in a non-experimental context through the joint probability distribution associated with the DAG.

A graphical condition called *back-door criterion* can be applied to a given causal graph to test if a subset of its nodes $\mathbf{S}$ is sufficient for identifying $P(Y|do(T = t))$ from observational data. A set of variables $\mathbf{S} \subseteq \mathbf{X}$ satisfies the back-door criterion relative to a graph G with node set $\mathbf{X}$, a treatment variable $T \in \mathbf{X}$ and an outcome variable $Y \in \mathbf{X}$ if:

1. no node in $\mathbf{S}$ is a descendant of T; and
2. $\mathbf{S}$ blocks all the paths between T and Y that contain a directed edge pointing towards T.

If the back-door criterion is satisfied by a set $\mathbf{S}$, then interventional quantities can be expressed through observational ones as follows:

$$P(y|do(T = t)) = \sum_{\mathbf{S}} P(y|t, \mathbf{s})P(\mathbf{s}) \tag{1.3}$$

The formula used to compute the interventional probability distribution of the outcome in Equation 1.3 is also known as adjusting for $\mathbf{S}$. Summary statistics of the interventional distributions can then be compared to compute the ATE.

Obtaining an adjustment set $\mathbf{S}$ through the back-door criterion also ensures that $\mathbf{S}$ satisfies the unconfoundedness condition for estimating the effect of T on Y. Therefore, performing a matching procedure (Imbens and Rubin 2015) by balancing the variable set $\mathbf{S}$, would ensure obtaining unbiased estimates of the ATE. This is an example of how Causal graphs can be used as guides for assessing and justifying the assumptions some PO methods require.

Suppose we are interested in estimating $P(y|do(T = t))$ given a causal model represented by the DAG G with node set $\{\mathbf{X}, T, Y\}$ in Figure 1.3 and a joint probability distribution $P(\mathbf{X}, T, Y)$. The knowledge of the DAG allows the application of the back-door criterion to select an adjustment set for causal effect estimation. The procedure reveals that adjusting for the set $\{X_3, X_4\}$ or $\{X_4, X_5\}$ ensures unbiased estimates of $P(y|do(T = t))$. Conversely, performing the adjustment procedure on a set $\mathbf{S} = \{X_4\}$ would produce biased estimates, since the set does not block all the back-door paths between X and Y.

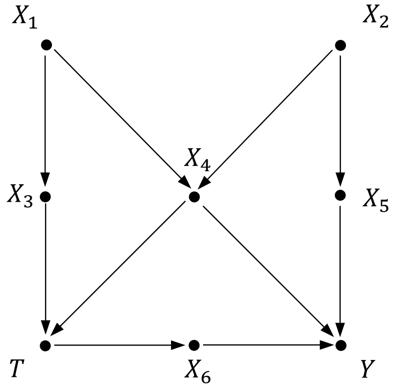

Figure 1.3. A DAG describing causal relations among a set of variables $\mathbf{X}$ a treatment T and an outcome Y

Let us now consider the graph in Figure 1.4. The DAG shows the presence of an unobserved or latent confounder U, which directly affects X_1 and T. The node U is denoted by a circle rather than a solid dot to indicate the variable is not observed. Even in the presence of unobserved variables, we can resort to the back-door criterion to assess if an adjustment set to estimate the effect of T on Y exists. In this scenario, we are particularly interested in checking if some of the sets that satisfy the back-door criterion are composed only by observed variables. Applying the criterion to the DAG reveals that $\{X_1\}$ and $\{X_2\}$ would be both valid adjustment sets and would thus ensure unconfoundedness.

Another example of a DAG with unobserved confounders is shown in Figure 1.5. The graph is very similar to the previous one, but now also X_2 is not observed, and the edge between X_1 and X_2 is oriented in the opposite direction. The node X_2 has thus have been replaced by U_2 to indicate it is a latent variable. In this situation, adjusting for $\{X_1\}$ would open the back-door path along the ordered tuple

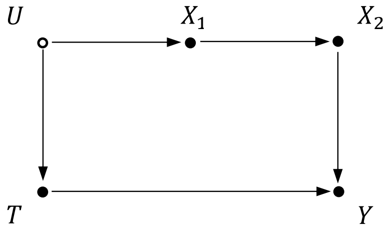

Figure 1.4. A DAG with unobserved confounders (1)

(T, U, X_1, X_2, Y), thus producing a biased estimate of the effect of T on Y. In this simple example, conditioning on the empty set provides instead unbiased estimates of the causal effect, since the colliding path over the ordered triplet (U_1, X_1, U_2) is blocked as long as we do not condition on X_1.

The bias introduced by conditioning on X_1 is also called $M - bias$, and it constitutes a solid motivating argument for employing causal graphs. Generally, the PO literature suggests to condition on all the observed pre-treatment variables in order to improve the quality of causal estimates (Imbens and Rubin 2015). However, in this scenario and similar ones, conditioning on the observed variables leads instead to worse causal estimates, and causal graphs provide a rule, namely the back-door criterion, to avoid this sort of bias. For a review on how conditioning can affect causal estimates, given different contexts represented by causal graphs, see Cinelli, Forney, and Pearl (2020).

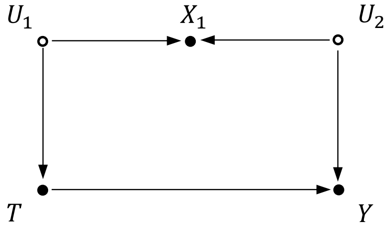

Figure 1.5. A DAG with unobserved confounders (2)

Front-door criterion and do-calculus

The back-door criterion is not the only strategy that can be employed to estimate causal effects from a causal graph. Pearl (1995) describes a specific graphical configuration that allows causal effect identification, even when back-door adjustment is not feasible. The condition is called *front-door criterion* and states that given a

DAG G with node set $\mathbf{X}$, a set $\mathbf{S} \subset \mathbf{X}$ satisfies the front-door criterion for the effect of T on Y, both belonging to $\mathbf{X}$, if:

1. $\mathbf{S}$ intercepts all directed paths from T to Y
2. all the back-door paths from T to $\mathbf{S}$ are blocked
3. all the back-door paths from $\mathbf{S}$ to Y are blocked by T

If a set $\mathbf{S}$ that satisfies the front-door criterion for the effect of T on Y exists and $P(t, s) > 0$, then the causal effect of T on Y can be computed with the formula

$$P(y|do(T = t)) = \sum_{s} P(s|t) \sum_{t'} P(y|t', s)P(t'). \tag{1.4}$$

Consider Figure 1.6 and assume we are interested in the effect of T on Y. We are also aware of the presence of an unobserved confounder U denoted in the DAG with a circle, instead of a solid dot. Variable U would satisfy the back-door criterion, but since it is unobserved, we cannot adjust for it to estimate the effect of T on Y. However, S satisfies the front-door criterion, and thus we can still identify the causal effect by applying Equation 1.4.

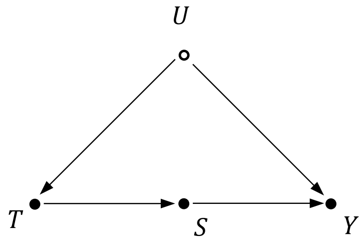

Figure 1.6. A DAG to illustrate the front-door criterion

Combined and iterative use of back-door and front-door criterion constitute the building block to identify causal effects on complex DAGs. Pearl (2000) describes a set of rules based on the two criteria, also called *do-calculus*, that allows expressing interventional distributions in terms of observational distributions only, in an automated way. The procedure has been proved to be sound and complete, meaning that an algorithmic iteration of the rules of do-calculus always return a solution for the identification of causal effects, if such solution exists (Pearl 1995; Tian and Pearl 2002; Huang and Valtorta 2012).

1.4 Causal discovery

Causal graphs are powerful models to describe the causal structure of a set of random variables. Moreover, they constitute a guide for selecting an identification strategy to estimate causal effects. However, the setting considered here always assumed a complete knowledge of the causal diagram.

Suppose we want to investigate the causal effect of a treatment variable T on an outcome variable Y from a dataset $D(\mathbf{X}, T, Y)$ where $\mathbf{X}$ is a set of other covariates. We also assume the existence of an unknown underlying causal model described by a DAG $G(V, E)$ and a joint probability distribution $P(V)$, from which $D(\mathbf{X}, T, Y)$ has been sampled. In order to obtain an unbiased estimate of $P(Y|do(T = t))$ we therefore study if it is possible to *learn* a causal graph from $D(\mathbf{X}, T, Y)$. In order to estimate the structure of the causal DAG, *structural learning algorithms* have been developed. These algorithms take a dataset as an input and, under a set of assumptions, recover a DAG and the associated joint probability distribution. This process is known as *causal discovery* (Spirtes et al. 2000).

Structural learning algorithms can be divided in three families: *constraint-based* algorithms, *score-based* algorithms and *hybrid algorithms.* Constraint-based algorithms learn the graph's structure via conditional independence statements emerging from data. They usually start with a complete graph, and then if two variables turn out to be marginally or conditionally independent, the edge connecting them is deleted. This procedure is repeated iteratively until a stopping criterion is satisfied. Score-based algorithms rely on a given score function that measures how well a certain DAG describes a dataset. These algorithms usually begin by computing the score of an initial graph. The diagram is then modified by introducing, deleting or reversing edges, and its score is computed again for each modification. The graph recording the best score at the end of the procedure is retained as the algorithm's output. Hybrid algorithms aim to exploit the advantages of score-based and constraint-based algorithms by merging them in a single procedure. Generally, they begin with a *restrict* phase where the parents of each node are selected through tests of conditional independence, similarly to what happens in constraint-based algorithms. The second phase is called *maximize* and consists in selecting a DAG in the restricted DAG family outlined by phase one by optimizing a given score function. Hybrid algorithms include the Max-Min Hill Climbing (Tsamardinos, Brown, and Aliferis 2006) and *H2PC* (Gasse, Aussem, and Elghazel 2014).

Once the graph is learnt, a joint probability distribution over the nodes of the graph can be obtained through maximum likelihood estimation. This phase usually involves computing maximum likelihood estimates subject to the independence constraints encoded in the graph. Estimates can be retrieved in the case of discrete variables or when dealing with continuous variables under the assumption of linearity (Spirtes et al. 2000).

The section will continue with a description of the assumptions that structural algorithms usually require. We will then explain how different algorithms work and show the functioning of two representative procedures.

1.4.1 Common assumptions and background knowledge

The assumptions of causal discovery algorithms usually focus on the relation between the causal graph and the distribution of the data employed to learn it. A usually required assumption is *faithfulness*. A graph G faithfully represents a dataset D, if all and only the conditional independence relations true in D are entailed by the Markov condition applied to G (Spirtes et al. 2000).

Another key assumption for learning algorithms is *causal sufficiency*. The assumption states that a given set of variables $\mathbf{X}$ is causally sufficient for a population if and only if in the population every common cause of any two or more variables belonging to $\mathbf{X}$ is in $\mathbf{X}$ or has the same value for all units in the population.

Implementing a constraint-based algorithm also requires making statistical decisions concerning how to assess conditional independence. Several tests can be employed to check if conditional independence holds, and violations of the assumptions required by the tests can generate unreliable independence statements. For a review of the implications of choosing a given independence test and what happens when the required assumptions do not hold, see Spirtes et al. (2000).

Structural learning algorithms are usually employed when information concerning the causal graph is not available. However, in practice it is common to deal with scenarios where the knowledge of the causal graph is partial. This incomplete knowledge can be introduced in structural learning procedures by imposing constraints on the structure of the obtained network. For example, if is known that a variable X_i cannot cause a second variable X_j, the directed edge that goes from X_i to X_j is forced to be absent. Note that this constraint does not imply the presence or absence of a directed edge going from X_j to X_i. Conversely, if background knowledge suggests that X_i affects X_j, a directed edge from X_i to X_j can be imposed. A consequence of including previous knowledge in the learning phase is that the graph is not entirely obtained through the information contained in the data. The constraints on the structure of the graph restrict the search space of the algorithms and often reduce both uncertainty and computational time.

1.4.2 Constraint-based algorithms

Constraint-based algorithms learn causal graphs from conditional independence relations contained in the data. They can take different kind of data as an input, including categorical and linear continuous variables: in the first case, the algorithm performs conditional independence tests on cell counts; in the latter, covariance matrices are used to test vanishing partial correlations. The obtained conditional independence statements, if possible, are then translated into graphical form according to the rules of d-separation. Constraint-based algorithm include the PC algorithm(C. Glymour, Spirtes, and Scheines 1991), the IG algorithm (Verma and Pearl 1990) and the most recent Grow-Shrink algorithm (Margaritis 2003). All the algorithms share the idea of learning a graph from the independence structure of the data but employ different heuristics. Constraint-based algorithms generally assume causal sufficiency, namely observing all the common causes of two or more variables in the model. This is a strong assumption, difficult to achieve in observational contexts. Some constraint-based algorithms have been proposed to deal with models where causal sufficiency does not hold. One of the most used is the fast causal inference (FCI) algorithm Spirtes et al. 2000. The algorithm is a variation of the PC algorithm and retrieves asymptotically correct causal structures in the presence of latent common causes, provided the observed distribution and the graph satisfy the faithfulness condition.

The PC-stable Algorithm

One of the most used algorithms in the constraint-based family is the *PC Algorithm*. The procedure begins with a complete undirected graph, in which edges are progressively deleted in order to generate a graph which is coherent with the conditional independence relations between variables. Faithfulness and causal sufficiency are assumed. A pseudocode of a recent variation of the algorithm, called *PC-stable* (Colombo and Maathuis 2014) is displayed in Algorithm 1. In the original PC algorithm the obtained graph could be affected by the ordering of the variables in the dataset used to learn the graph. In the new version, instead, the ordering does not affect the results, thus the name PC-stable. The procedure begins by learning a graph containing only undirected edges from conditional independence statements retrieved from the dataset. Then the orientation of the edges is estimated according to a set of graphical rules. In the pseudocode we will denote directed and undirected edges between to nodes X_i and X_j, respectively with the notation $X_i \rightarrow X_j$ and $X_i - X_j$. Moreover we will use $adj(X_i)$ to denote the set composed by the nodes adjacent to X_i and $\mathbf{X}\backslash\{X_i\}$ to indicate the variable set $\mathbf{X}$ excluding variable X_i.

Algorithm 1: PC-stable

```
   Input: A sample $D = (\mathbf{X})$ from a set of random variables $\mathbf{X} = \{X_1, ..., X_N\}$ and a chosen
          statistical test of conditional independence
   Output: A family of Markov-equivalent DAGs
 1 Form a complete undirected graph $G$ with vertex set $\{X_1, ..., X_N\}$;
 2 Set $l = -1$;
 3 repeat
 4     $l = l + 1$;
 5     forall vertices $X_i$ in $G$ do
 6         Set $a(X_i) = adj(G, X_i)$
 7     end
 8     repeat
 9         select a (new) adjacent pair of nodes $(X_i, X_j), i \neq j$ in $G$ such that $|a(X_i)\backslash X_j| \geq l$;
10         repeat
11             Choose a (new) set $\mathbf{S} \subseteq a(X_i)\backslash\{X_j\}$ of size $l$;
12             if the statistical test reveals that $X_i$ is conditionally independent from $X_j$ given $\mathbf{S}$
                 then
13                 delete the edge connecting the pair $(X_i, X_j)$ from $G$;
14                 set $\mathbf{S}_{X_i X_j} = \mathbf{S}$, denoting the set that separates $X_i$ and $X_j$
15             end
16         until $X_i$ and $X_j$ are no longer adjacent in $G$ or all possible subsets $\mathbf{S}$ of size $l$ have been
             considered;
17     until all pairs of adjacent nodes $(X_i, X_j), i \neq j$ in $G$ such that $|a(X_i)\backslash\{X_j\}| \geq l$ have been
         considered;
18 until all pairs of adjacent nodes $(X_i, X_j)$ in $G$ satisfy $|a(X_i)\backslash\{X_j\}| \leq l$;
19 foreach triplet $\{X_i, X_k, X_j\}$ such that $X_i$ is adjacent to $X_k$, the latter is adjacent to $X_j$, but the
     pair $\{X_i, X_j\}$ is not adjacent to $X_j$, if $X_k \notin \mathbf{S}_{X_i,X_j}$ do
20     orient $X_i - X_k - X_j$ with the colliding configuration $X_i \rightarrow X_k \leftarrow X_j$.
21 end
22 Set more arc directions by repeated application the following rules:
23 if $X_i$ is adjacent to $X_j$ and there is a directed edge from $X_i$ to $X_j$ then
24     replace $X_i$ - $X_j$ with $X_i \rightarrow X_j$
25 end
26 if there are two paths $X_i - X_k \rightarrow X_j$ and $X_i - X_l \rightarrow X_j$ and $X_k$ is not adjacent to $X_l$ and there
     is a directed edge from $X_i$ to $X_j$ then
27     replace $X_j$ - $X_k$ with $X_j \rightarrow X_k$
28 end
29 if $X_i$ and $X_k$ are not adjacent but $X_i \rightarrow X_j$ and $X_j$ - $X_k$ then
30     replace $X_j - X_k$ with $X_j \rightarrow X_k$
31 end
```

Given a dataset $D = (\mathbf{X})$ describing a set of random variables $\mathbf{X} = \{X_1, ..., X_N\}$, the PC-stable algorithm begins by forming a complete undirected graph G over $\mathbf{X}$. Then, step 5 stores the adjacency sets $adj(G, X_i)$ for each node X_i according to the current structure of G. Given an index l which start from 0 and increase at each iteration, the procedure checks if a set $\mathbf{S}$ of size l, that d-separates two nodes X_i and X_j exists. Note that $\mathbf{S}$ must be formed by nodes belonging to $adj(G, X_i)$ obtained in step 5 and that the size of $adj(G, X_i)\backslash X_j$ must be greater or equal than l. If the procedure finds a set $\mathbf{S}$ of size l that makes X_i and X_j conditionally independent, the edge between them is deleted from G and $\mathbf{S}$ is retained. The procedure is repeated for every node pair (X_i, X_j) and for every possible size l $\mathbf{S}$ associated to it, until an $\mathbf{S}$ that ensures d-separation is found or every size l $\mathbf{S}$ has been explored. The algorithm then increases l by a unit and repeat the procedure from step 5, until every pair of adjacent nodes (X_i, X_j) in G satisfies $|a(X_i)\backslash\{X_j\}| \leq l$. In other words, at each iteration, the structure of G is updated by removing edges between conditional independent variables. Once the undirected graph is obtained, steps 19-31 orient the edges according to specific edge configurations. The rules dictated by the algorithm ensure that cycles are not generated and avoid the creation of a new colliding configuration that would modify the conditional independence relations.

The output of the PC-stable algorithm is a *completed partially DAG*(CPDAG), a DAG where some of the edges are undirected. This kind of graph is used to represent a family of independence-equivalent DAGs. Regardless of how the undirected edges of the graph are oriented, the colliding configurations remain the same, thus ensuring that the all the DAGs associated to a CPDAG encode the same conditional independencies. The output of the PC-stable algorithm is therefore coherent with the objective of translating the conditional independences contained in the data into graphical form. Moreover, it has been proven that if the assumptions hold, the results provided by the algorithm are sound and complete (Colombo and Maathuis 2014).

1.4.3 Score-based algorithms

Score-based algorithms aim at recovering the graph structure from data by optimizing a score function. The score function evaluates the goodness of fit of the graph with respect to the learning data. Common choices for the score function are the likelihood function or the Bayesian Information Criterion (BIC), for a comprehensive review of the available score functions see Koller and Friedman (2009) . Generally, this kind of algorithm explores several graph structures and assigns a score to each of them; at the end of the procedure, the graph with the maximal score is retained. Score-based algorithms usually assume faithfulness as well as causal sufficiency. Algorithms belonging to this family include the *greedy search*, the *simulated annealing* and *genetic algorithms* (Russell and Norvig 2009).

Greedy Search algorithm

One of the most used score-based algorithms is the *greedy search* and its steps are shown in the pseudocode of Algorithm 2. The procedure iteratively modifies the edges of an initial DAG, computes the score of each graph and retains the best-scoring structure. When the score does not increase with an iteration, the obtained graph is provided as the algorithm's output.

Algorithm 2: Greedy Search

```
   Input: A sample $D = (\mathbf{X})$ from a set of random variables $\mathbf{X} = \{X_1, ..., X_N\}$ a score function
          $\mathcal{F}(G, D)$
   Output: A DAG
 1 Form an empty graph $G$ with vertex set $\{X_1, ..., X_N\}$;
 2 Calculate the score of $G$ given $D$, $S_G = \mathcal{F}(G, D)$ ;
 3 Set $S_{max} = S_G$ ;
 4 Set $G_{max} = G$ ;
 5 repeat
 6     foreach possible edge addition, removal or inversion in $G_{max}$ that produces a modified DAG
         $G^\star$ do
 7         compute $S_{G^\star} = \mathcal{F}(G^\star, D)$;
 8         if $S_{G^\star} > S_{max}$ and $S_{G^\star} > S_G$ then
 9             set $G = G^\star$ and $S_G = S_{G^\star}$
10         end
11     end
12     if $S_G > S_{max}$ then
13         set $S_{max} = S_G$ and $G_{max} = G$
14     end
15 until $S_{max}$ of current iteration is smaller then $S_{max}$ of previous iteration;
```

Given a dataset $D = (\mathbf{X})$ and a score function $\mathcal{F}(G, D)$, the algorithm first two steps consist in computing the score of an initial, usually empty, graph G with vertex set $\mathbf{X}$. Next, the score of the graph is set as the maximal score S_{max} and the initial graph G is set as the best-scoring DAG G_{max}. In step 6, the best-scoring DAG is modified by deleting, adding or inverting an edge, thus generating a new DAG $G^\star$. The score of $G^\star$ is computed, and if it is greater than the best score of the iteration S_G and greater than the absolute best score S_{max} then $G^\star$ becomes the new best score of the iteration S_G. All the possible modifications to G_{max} are explored this way, and if the best-obtained score of the iteration is greater than the best absolute score, then the latter is set to the current S_G and G_{max} is set equal to the current G. The procedure is then repeated from step 6 for the new G_{max}. The algorithm stops when applying all the possible modifications to the DAG G_{max}, obtained in the previous iteration, does not generate an increased S_{max}. In this case, G_{max} constitutes the output of the algorithm.

1.4.4 Causal discovery and potential outcomes

We have already shown how PO methods can benefit from specifying a causal graph to outline causal relations between variables. If the causal knowledge is available, drawing a causal graph can help assess unconfoundedness in a high dimensional context, using a graphical condition called the back-door criterion.

Causal discovery methods constitute an additional resource if we are interested in estimating causal effects with PO methods, but the knowledge of the causal graph is partial or absent. Suppose we want to estimate the effect of treatment T on an outcome Y given a dataset $D(\mathbf{X}, T, Y)$, where $\mathbf{X}$ are additional random variables, that could directly or indirectly affect T and Y. In addition, let us assume that the available subject matter knowledge concerning the variable causal structure is very limited and thus does not allow drawing a causal graph. In order to estimate causal effects with a PO method such as matching, we have first to assess if unconfoundedness holds. However, since the causal graph over $\{\mathbf{X}, T, Y\}$ is unknown, we cannot directly select an adjustment set $\mathbf{S}$ that satisfies the back-door criterion.

Causal discovery provides a solution to this scenario. If we cannot exclude the absence of unobserved common causes, we can learn the graph from $D(\mathbf{X}, T, Y)$ employing an algorithm that only requires the faithfulness assumption, such as the FCI algorithm. The algorithm's output can be then used to assess which PO identification strategy is adequate to estimate the causal effect of T on Y. If instead, it is reasonable to assume both causal sufficiency and faithfulness, we can opt for an algorithm such as the greedy search or PC-stable. In both cases, we know that if the assumptions hold, the obtained causal structures are asymptotically correct, and a sufficient adjustment set can be selected by applying the back-door criterion. The adjustment set can then be used to derive the interventional distribution through the adjustment formula, or directly estimate the ATE with a method of choice, such as regression, matching or inverse probability weighting.

Alternatively, learning the graph from data could reveal or confirm if a specific PO identification strategy is feasible. Assume that applying a structural learning algorithm on a given dataset generates the DAG in Figure 1.7.

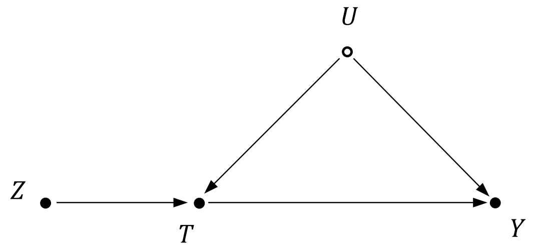

Figure 1.7. Instrumental variable DAG

If we are interested in the effect of T on Y, we cannot directly estimate causal effects because of the presence of the unobserved confounder U, and no observed adjustment set that satisfies the back-door criterion. However, the graph configuration reveals that variable Z satisfies the exclusion restriction assumption of instrumental variables described in Section 1.2. This means that we can employ an IV strategy to achieve causal effect identification. Also in this case, the assumptions required by PO methods are made transparent by causal graph implementation. In this particular example, those assumptions are also strengthened by the structural learning procedure that allows exclusion restrictions to be derived directly from the data.

1.5 Discussion

Estimating causal effects is a central subject for biomedical and social sciences. However, investigating causal claims is an ambitious objective, especially when dealing with observational data. The most affirmed causality frameworks are Potential Outcomes and Causal Graphs.

The two approaches are often contrasted to evaluate which one is most effective. PO methods offer efficient ad hoc solutions to specific causal problems. However, their assumptions are difficult to assess, especially as the number of variables increases. On the other hand, causal graphs allow the formalization of complex causal problems in a generalized way. Nevertheless, their high generality can sometimes be perceived as a distance from real empirical problems and incapacity of including context-specific restrictions in the model.

This paper described how the two frameworks could be implemented together in an integrated approach. Causal graphs can be used as a guide to evaluating which PO method can be implemented and if its assumptions hold. The graph can be outlined directly if the causal structure is entirely known or learned from data if the causal knowledge is partial or absent. This versatility guarantees coverage of most empirical problems. The results of PO methods are thus strengthened by causal graphs, since assumptions such as unconfoundedness and exclusion restrictions can be directly assessed from the structure of the DAG. At the same time causal graphs can benefit from all the context-specific identification strategies provided by the literature of potential outcomes. Combining the two methodologies thus results in an effective synergic approach that enhances both frameworks' peculiar characteristics.

Chapter 2

Evaluating the effect of home-based working on firms' expected revenues during the pandemic

Abstract

Covid-19 generated an unprecedented shock on the Italian economy, which severely affected firm performance. This work focuses on estimating the causal effect of implementing home-based working (HBW) after the pandemic outbreak on firms' expected revenues. The analysis uses a unique firm-level dataset, which captures a rich set of features before and after the spread of the virus. Causal effect estimation is performed implementing an integrated approach that merges Causal Graphs and Potential Outcomes frameworks. At first, the dataset is used to learn a causal diagram that encodes theory-based assumptions and information contained in the data. An adjustment set is then selected by applying the back-door criterion on the obtained graph. Lastly, causal estimates are computed with full matching, using the chosen adjustment set to ensure unconfoundedness. The results confirm the presence of a positive effect of the implementation of HBW on expected revenues. The treatment seems to be particularly effective in providing revenue stability and mitigating of losses. The results are consistent with the fact that HBW equips firms with greater flexibility and helps contain productivity decreases in Covid times.

2.1 Introduction

The outbreak of Covid-19 in March 2020 had unprecedented consequences on the Italian economy. As the virus spread, consumer spending dropped, and lockdown policies forced many firms to temporarily cease their activity, thus generating both a demand and a supply shock. As soon as the economic consequences of the covid outbreak became clear, firms tried to do everything possible to minimize losses.

This work focuses on the implementation of home-based working (HBW), one of the key firms' countermeasures to the pandemic. In particular, given the firms' characteristics, the analysis evaluates the effect of HBW on expected revenues by comparing firms who implemented home working with those who did not.

The implications of switching to HBW have been thoroughly studied over the past years and its related literature has spiked in covid times. The benefits of home working on employees performance have been studied in Bloom et al. (2015). The research points out that when staying at home, employees adopted longer working shifts and showed increased productivity. On the other hand, evidence from workers who switched to home working in covid times suggests that being far from the workplace for a prolonged period can negatively affect mental health (Felstead and Reuschke 2020). Among the possible effects of working from home, there are also income inequalities, as stated in Bonacini, Gallo, and Scicchitano (2021). The authors warn about the risks of amplifying pre-existing inequalities by favouring male, older, high-educated and high-paid employees.

Bartik et al. (2020) use firm-level surveys to investigate the spread of home-based working during the pandemic. The authors find out that industries with more educated workers are associated with a higher rate of remote working and perceive a lower productivity loss associated with this kind of work. In addition, about 40% of interviewed firms declare that at least 40% of their workers that switched to homeworking will continue doing so even after the crisis, and this represents a strong indicator of the persistence of the phenomenon. These results are also confirmed in Bick, Blandin, and Mertens (2021) and Barrero, Bloom, and Davis (2021).

This work contributes to the fast-growing literature of home-based working by evaluating the effect of implementing home working in covid times on future expected firm revenues. A rigorous causal evaluation of this kind seems to be missing in the literature and could provide a quantifiable measure of the impact of enabling employees to work from home.

A picture of the condition of firms just before and after the outbreak is needed to perform the analysis. A private firm-level dataset has been employed to answer this need. The dataset originates from two different surveys over the same group of firms: one was conducted just before the covid outbreak and one a few weeks after. Both surveys have been provided by MET *Monitoraggio Economia e Territorio*, a private research centre that regularly conducts one of the most comprehensive private surveys on Italian manufacturing and production companies. The pre-covid survey covers many firms' characteristics, including financial and strategic components. The second survey has been conducted to capture the immediate effect of the shock on firms' organizations and the change in their future expectations. Having such a rich dataset is an added value compared to similar studies and is crucial to obtaining a comprehensive representation of the problem and making causal estimates more credible.

One of the possible frameworks to investigate causal claims when dealing with observational data are Causal Graphs (Pearl 1995). In this approach, the causal model comprises a joint probability distribution and a graph, which entails the relations between variables in the form of nodes and edges. An alternative and widely used approach in economics is the Potential Outcomes framework (Rubin 2005). The methods belonging to this approach aim at checking if the main features of randomized experiments still hold or can be emulated in some particular cases of observational studies (Imbens 2020). The main concern for both frameworks is confounding: the situation where at least one variable has a direct causal effect on both the treatment and the outcome. This kind of configuration can generate biased estimates unless adequately accounted for.

This work integrates the two approaches: causal graphs are used to cope with uncertainty in the causal structure, and then causal effects are estimated through potential outcomes methods. In the first step, a structural learning algorithm estimates a causal graph from data. The obtained model provides a clear picture of the interactions between variables and allows a straightforward identification of confounders. The generalized back-door criterion (Perković et al. 2015) is then used to select a sufficient set of variables for confounding adjustment, according to the graph's structure. Lastly, the causal effect is estimated using full matching on the chosen adjustment set.

The analysis results suggest that home working affected performance during the pandemic, generating a substantial increase in the expected revenues of firms that implemented it relatively to those who did not. The proposed integrated methodology reconstructs a credible causal graph and provides a rigorous framework for unbiased causal effect estimation. In addition, the validity of causal estimates is strengthened by the unique dataset that captures a rich set of firm dimensions, including implemented strategies, workforce characteristics and financial structure, thus ensuring a comprehensive representation of the problem.

The paper is organized as follows. In Section 2.2, we will describe the dataset that has been used for the analysis, what are its sources and why it has been chosen. In section 2.3 both causal graphs and full matching are explained in detail, their assumptions are outlined, and the strengths of opting for a combined approach are highlighted. The obtained causal graph, the selected adjustment set, matching results and the estimated causal effect are presented in Section 2.4. Finally, results and future work are discussed in Section 2.5.

2.2 Data

2.2.1 Data source

The analysis uses a unique firm-level dataset provided by MET, a research centre based in Rome, which conducts one of the most comprehensive surveys on the Italian manufacturing and production service sectors. The dataset originates from merging two different MET surveys over the same panel of firms. The same data source has already been employed to study the economic effects of the Covid-19 shock in E. Brancati and R. Brancati (2020) and Balduzzi et al. (2020).

The first survey is the 2019 wave of the MET survey on the italian industrial system. The questionnaire covers a vast group of firm features such as structure, performance and strategies. Almost 24000 firms were interviewed according to their size, sector, and area to obtain a representative sample of the Italian manufacturing and production services population. The administration began at the end of 2019 and was completed in late January 2020, right before the spread of the pandemic in Italy.

Reaction to Covid-19 was then measured with another questionnaire between March 24 and April 7, 2020, administered to the 24000 respondents of the 2019 MET survey. The Italian government imposed restrictive measures for firms on March 8, thus leaving time for the second survey respondents to adapt to such change. The time window for completing the survey has been restricted to two weeks to avoid as much as possible that variations in the regulations or the spread of the disease would generate heterogeneous answers. The Covid-19 survey is divided in three parts:

1. A section that replicates questions from the MET 2019 survey concerning expected sales and future R&D plans. This first block allows a direct comparison between the answers of the two surveys to investigate the effects of the pandemic.

2. A section that complements the first one, asking directly how Covid-19 will affect firms' future operations and performance. This part is intended to study additional consequences of the pandemic which were impossible to study from a comparison with the MET 2019 survey.

3. A section with questions related to the beahviour and needs of firms during the pandemic. This last part provides information regarding how firms perceive Covid-19-related risks, the reaction strategies they implemented and the public policies they demand to aptly face the pandemic.

The exceptional timing thus produces two snapshots of the same group of firms, just before and after the spread of the pandemic. The answers to both surveys have been merged to obtain a final dataset of 7800 respondents.

2.2.2 Description of the dataset

Fifteen variables have been selected out of the set of data. No weighting scheme has been used on the dataset, and therefore the results of the analysis are tied to the population represented by the interviewed sample. An overview of the composition of the sample can be found in Table 2.1, which contains descriptive statistics of all the selected variables.

The treatment variable originates from the post-covid questionnaire and is a binary variable defining if a firm has implemented home-based working for a portion of their employees right after the lockdown policies introduction. The idea behind this treatment is that it primarily represents an indicator of the preparedness of a firm to switch to HBW. It is assumed that having at least a portion of the employees working from home has been convenient for most firms in covid times. Therefore the firms which declared that they did not resort to HBW at all after the outbreak are assumed to be those who were not able to face the transition. The treatment thus describes the capability of a firm to implement HBW. The feasibility of this strategic change could depend on several factors, such as employee education, manager education and the degree of digital literacy in the firm. A rich set of variables will be included in the model to investigate these interactions.

The outcome variable describes post-covid expectations towards future variation in revenues with respect to past revenues. The variable derives from the post-covid survey and can take four different modalities, which identify increase, stability, decrease or strong decrease in expected revenues. Covid-19 and related lockdown policies strongly affected expectations regarding future revenues. The magnitude of this phenomenon can be appreciated by comparing the outcome variable with the same variable measured just before the outbreak. Pre and post-covid survey present several similarities in their structure, which allow a direct variable comparison and thus provide an idea of the effect of the pandemic. The change in the distribution of expected revenues is represented in Figure 2.1.

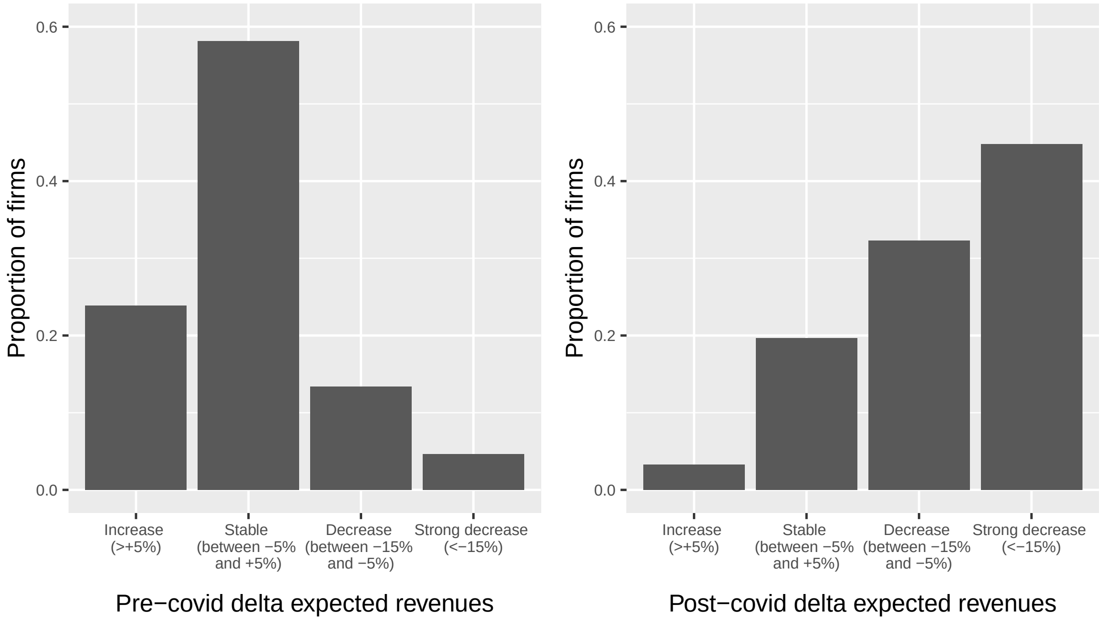

Figure 2.1. Distribution of Δ expected revenues before and after the covid outbreak

Firms are hugely affected by the shock, which dramatically impacts expectations. Before March 2020, most of the interviewed firms (58%) imagined a stable revenue flow in the following two years; almost a quarter (24%) expected an increase and the remaining portion (18%) was preparing to withstand a decrease. After the spread of the virus, the distribution changed radically, and more than three-quarters of respondents (77%) declared to believe they would have registered a decrease in revenues in the next two years. This radical shift towards negative expectations provides a provisory yet powerful picture of the magnitude of the shock.

The other variables included in the model can be divided into groups based on their type. These groups will also serve as a guide to build the causal diagram in Section 2.4.1. Variable names will be denoted with the *italic* font.

– The first group of covariates defines the firm's structure and is constituted by firm *size (n. of employees)*, *geographical area*, *economic sector* and *manager education*. These typifying characteristics are usually correlated with each other and could highly affect firm behaviour and performance.

– The second ensemble is more heterogeneous and is formed by other characteristics measured before the spread of the virus, such as implemented strategies, past performance and financial constraints. This group includes information regarding past research and development activity (R&D), innovation, export, employees and revenues variation, credit rationing and level of digitalization. Variables belonging to this group are derived from survey questions that investigate firm behaviour in the previous three years. For example *R&D activity and innovation* is a binary variable which takes value 1 if one of the two strategies has been adopted by the firm in that time interval and 0 otherwise. Employees and revenues variation are also computed with respect to the same reference period. The level of digitalization is captured by the variable *Digital literacy* which takes value 1 if the firm trained its employees to improve their IT skills, adopted high-tech equipment or made other investments in ICT and 0 otherwise.

– The third group consists of a single variable describing the pre-covid expectations concerning future revenues variation. Including it in the model allows us to account for previous expectations when estimating treatment effects.

– The fourth and last group contains two variables that are not affected by the treatment but refer to a time window that follows the pre-covid survey. The first one is *confirmed covid infections*, a variable that describes the number of confirmed covid infections at the province level and acts as a proxy of the geographical heterogeneity of the spread of the virus. More precisely, the variable represents the number of confirmed infections released by the Italian Department of Civil Protection in the province of the firm, the day before answering the questionnaire. This inclusion tries to capture both the physical and the psychological effects of the virus. The number of detected infections was announced on the media daily and it was the most followed indicator of the propagation of the virus. The other variable is *essential business sector*. Lockdown policies implemented by the government at the beginning of March forced many firms to cease their activity. In particular, depending on their business sector, the regulation allowed only firms belonging to ”essentials” business areas to stay open, whereas the rest were forced to close. Moreover, some of the essential firms decided to shut down even if allowed to remain open. The variable *essential business sector* was added to the dataset in order to take into account the effects of these measures. The economic sector has been used to distinguish essential from non-essential firms and integrated with the answers of the post-covid survey to account for firms that voluntarily shut down even if belonging to essential sectors.

2.3 Methodological background

In this section, the main methodological elements used for causal effect estimation are described. We begin by explaining the basic concepts of causal graph theory and how a causal diagram can be learned from data. Next, we proceed with how the graph can be used to obtain a sufficient adjustment set for unbiased causal effect estimation. Finally, the potential outcome framework is introduced, and full matching is explained in detail.

2.3.1 Causal graphs

Causal graphs are models providing a clear representation of causal problems and a set of tools to derive causal estimates. The theoretical framework is described in Pearl (1995) and shares elements of similarity with path diagrams found in the work of Wright (1921a).

Graph terminology

A graph $G = (\mathbf{X}, \mathbf{E})$ is a collection of nodes $\mathbf{X}$ and edges $\mathbf{E}$. When an edge goes out from a node into another is called a directed edge, if there is no such orientation the edge is undirected. A graph that contains only directed edges is a directed graph. Given a graph G with ensemble of nodes $\mathbf{X}$ and two nodes X_i and X_j belonging to $\mathbf{X}$, any sequence of edges which connects X_i and X_j, regardless of their direction, is called a path. If every edge of the path is directed and has the same orientation along the path, then it is called a directed path. A directed path which begins and

Table 2.1. Descriptive summary of the variables included in the model

Variable	Level description	Proportion
Home-based working	0 Not implemented	0.64
	1 Implemented	0.36
Post-covid Δexpected revenues	1 Increase (>+5%)	0.03
	2 Stable (between -5% and +5%)	0.20
	3 Decrease (between -15% and -5%)	0.32
	4 Strong decrease (<-15%)	0.45
Size (n. of employees)	1 1-9	0.51
	2 10-49	0.33
	3 50-249	0.13
	4 >250	0.03
Geographical area	1 North-West	0.25
	2 North-East	0.27
	3 Center	0.24
	4 South and islands	0.24
Business sector	1 Food and beverage	0.07
	2 Textile	0.06
	3 Wood industry	0.05
	4 Paper industry	0.05
	5 Plastic industry	0.06
	6 Metal industry	0.09
	7 transport equipment	0.02
	8 Machinery ed equipment	0.11
	9 Electrical and optical equipment	0.05
	10 Production and distribution of utilities, mineral extraction	0.07
	11 Services for the manifacturing industry	0.37
Manager education	0 <20% of managers achieved a degree	0.67
	1 >20% of managers achieved a degree	0.33
Innovation, R&D	0 No	0.40
	1 Yes	0.60
Credit rationing	0 No	0.92
	1 Yes	0.08
Export	0 No	0.70
	1 Yes	0.30
Delta number of employees	1 Decrease	0.22
	2 Stable	0.45
	3 Increase	0.32
Digital literacy	0 No	0.57
	1 Yes	0.43
Past delta revenues	1 Increase (>+5%)	0.34
	2 Stable (between -5% and +5%)	0.46
	3 Decrease (between -15% and -5%)	0.14
	4 Strong decrease (<-15%)	0.06
Pre-covid delta expcted revenues	1 Increase (>+5%)	0.24
	2 Stable (between -5% and +5%)	0.58
	3 Decrease (between -15% and -5%)	0.13
	4 Strong decrease (<-15%)	0.05
Confirmed covid infections	0 Low	0.25
	1 Medium	0.25
	2 High	0.25
	3 Very high	0.25
Essential business sector	0 No	0.26
	1 Yes	0.74

ends with the same node is a cycle. If a directed graph does not contain cycles then it is a directed acyclic graph (DAG). In the context of this work, the nodes of the DAG represent random variables and edges describe the causal relations between them. To introduce some further terminology, consider the simple graph in Figure 2.2(a). The model describes a treatment variable T which is the only cause of an outcome variable Y. The causal relation is represented through the directed edge that links the two nodes denoting the variables. When two nodes are connected by an edge they are called adjacent. Since the directed edge goes from T to Y, then T is a parent node of Y and Y is a child of T. If we assume that T affects Y only through a third variable X which acts as a mediator, the model could be instead described by the graph in Figure 2.2(b). This particular configuration where every node has at most one parent and one child, is called a *chain*. The edges of the chain also constitute a directed path which connects T and Y. For a given DAG G, the structure of the graph allows factorizing the joint probability distribution of its nodes $(X_1, ..., X_n)$ as follows

$$P(X_1, ..., X_n) = \prod_i P(X_i|pa_i) \tag{2.1}$$

where pa_i is the set of the parents of X_i according to the graph. In the case of the graph in Figure 2.2(b) we would have $P(T, X, Y) = P(T)P(X|T)P(Y|X)$.

Let us consider a slightly more complex model in Figure 2.2(c). A variable Z causes T, the treatment directly causes Y, and a fourth variable X causes both T and Y. Assuming that we are interested in the effect of T on Y and that we observe Z but not X, the DAG constitutes a graphical representation of the instrumental variable (IV) context. The configuration between X,T,Y, is called confounding and, unless accounted for, produces a bias in the estimates of the causal effect of T on Y. When instead two non-adjacent nodes point to a third node, the latter is called a collider on the path formed by the triplet of nodes. In the IV graph, the treatment variable T is a *collider* on the path $(Z \rightarrow T \leftarrow X)$, where arrows denote directed edges.

Conditional independence statements are encoded in a DAG through the rules of $d-separation$. Two nodes are d-separated, and thus independent, if every path between them is blocked. Consider a DAG G with nodes $\mathbf{X}$, a pair of nodes X_i and X_j belonging to $\mathbf{X}$ and a set of nodes $\mathbf{S} \subset \mathbf{X}$. A path p which connects X_i to X_j is blocked, conditioning on $\mathbf{S}$, if

1. p contains a chain of nodes $n_1 \rightarrow n_2 \rightarrow n_3$ or a configuration of the kind $n_1 \rightarrow n_2 \leftarrow n_3$, such that the node n_2 is in $\mathbf{S}$, or

2. p contains a collider $n_1 \leftarrow n_2 \rightarrow n_3$ such that the node n_2 is not in $\mathbf{S}$ and that

no descendant of n_2 is in $\mathbf{S}$.

This also implies that when all the paths between X_i and X_j contain colliders and $\mathbf{S} = \emptyset$, X_i and X_j are d-separated. In this case X_i and X_j are independent without conditioning on $\mathbf{S}$ and are thus said to be marginally independent. On the other hand, if two nodes are d-separated only by conditioning on a non-empty set $\mathbf{S}$, they are said to be conditionally independent.

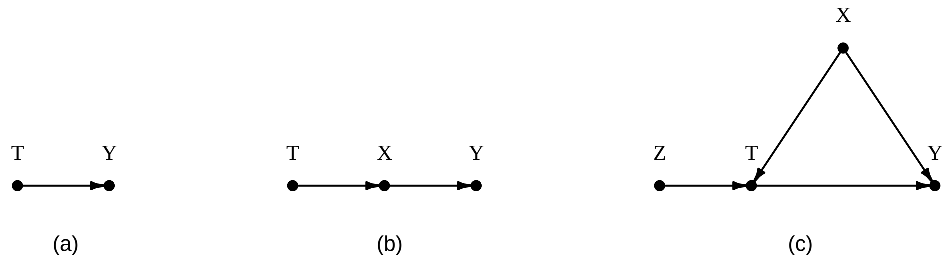

Figure 2.2. Simple DAGs

Theory of intervention

A DAG and the joint probability distribution associated to its nodes is also known as a Bayesian Network (BN). BNs can be provided with an inference engine that allows to investigate how the model reacts if some evidence is introduced, through a method called *what-if* analysis. This procedure performs an update of the probability distribution given that we observe one or more variables to assume a specific state (Kjaerulff and Madsen 2008). However, to answer causal queries, we are interested in studying how the model would react to an *intervention* on one or more variables. The difference between observing and forcing a variable to take a particular state is a substantial one and requires some further discussion. Pearl (2009a) introduces the notation $do(X_i = x_i)$ to denote that a variable X_i is set to the value x_i through an intervention. The quantity $P(X_j|do(X_i = x_i))$ represents then the distribution of X_j given that X_i is forced to take value x_i, while $P(X_j|X_i = x_i)$ describes the distribution of X_j given that we observe X_i take value x_i. Interventional quantities can be used to estimate causal effects, that can be expressed as comparison of interventional distribution summary statistics. A common way to represent causal effects is the *average treatment effect* (ATE) (Imbens and Rubin 2015). If we consider the effect of a binary variable X_i on a variable X_j, the ATE can be computed as

$$ATE = E[X_j|do(X_i = 1)] - E[X_j|do(X_i = 0)] \qquad \text{or} \qquad ATE = \frac{E[X_j|do(X_i = 1)]}{E[X_j|do(X_i = 0)]}$$

However, in observational studies, the interventional distribution $P(X_j|do(X_i = x_i))$ is not directly measured. In order to express this distribution through observational quantities, Pearl (1995) introduces the back-door criterion. In particular, given a graph G with ensemble of nodes $\mathbf{X}$, a treatment T and an outcome Y belonging to $\mathbf{X}$, if a set $\mathbf{S} \subset \mathbf{X}$ satisfies the following assumptions

1. No $s \in \mathbf{S}$ is a descendant of T
2. $\mathbf{S}$ blocks every path between T and Y that contains an edge pointing to T

then interventional distributions can be expressed in observational terms:

$$\begin{aligned}&P(Y|do(T = t)) = \\ &= \sum_{\boldsymbol{S}} P(Y|\mathbf{S} = \mathbf{s}; do(T = t))P(\mathbf{S} = \mathbf{s}; do(T = t)) \\ &= \sum_{\boldsymbol{S}} P(Y|\mathbf{S} = \mathbf{s}; T = t)P(\mathbf{S} = \mathbf{s})\end{aligned} \tag{2.2}$$

A set $\mathbf{S}$ that satisfies the assumptions is called a *sufficient adjustment set* (Greenland, Pearl, and Robins 1999). The paths between T and Y with an edge pointing to T, as mentioned in Assumption 2, are called *back-door paths.* These paths are carriers of spurious associations between outcome, and treatment and their individuation is crucial in order to obtain unbiased causal estimates. The bias introduced by back-door paths is also referred to as confounding bias and is one of the main concerns when drawing causal conclusions from observational data. However, as shown in (2.2), given a known causal graph, an adjustment set obtained through the back-door criterion allows the calculation of unbiased interventional distributions. The selected set can be used for adjustment through various methods, including matching, weighting, regression adjustment or doubly robust methods (Abadie and Cattaneo 2018).

Graph learning

We have until now described the main features of causal graphs and how they can be used to answer causal queries. However, the proposed methods strongly rely on the structure of the graph, which is often partially or entirely unknown when dealing with real problems. In this case, the causal graph can be recovered from a dataset containing the variables of interest. Extracting conditional independence statements from the data and encoding them into a DAG, usually requires the following conditions (Pearl 2000):

Faithfulness condition. All the independence assumptions that can be read-off from the graph by d-separation rules are exactly the same as those in the population that generated the DAG. This assumption, also known as stability, ensures that independence statements encoded in the graph are consistent with the factorization in 2.1.

Causal sufficiency. There are no unobserved variables that would invalidate the factorization in 2.1.

If the conditions hold, causal structural learning can be carried out using one of the algorithms that have been developed to perform this task. For a comprehensive review, see Stuart Russell and Norvig (2002). Given a dataset D containing n observations, the algorithms output a graph G following one of three possible approaches: *constraint-based*, *score-based* and *hybrid.* Constraint-based algorithms retrieve the structure of the graph by performing a sequence of conditional independence tests. Those tests produce a set of conditional independence statements, which are then encoded in a DAG G following the rules of d-separation. *Score-based* algorithms instead select the DAG which maximizes a score reflecting its goodness of fit. *Hybrid algorithms* are combinations of the two approaches: they use conditional independence tests in a first step to restrict the space of possible graph structures and in a second phase, select the DAG which maximizes a given network score.

2.3.2 Potential outcomes

Potential outcomes (Splawa-Neyman, Dabrowska, and Speed 1990; Rubin 2005) are an alternative framework to deal with causality. This approach is widely used in economics and allows estimating causal effects from experiments and some specific observational contexts. The framework's name originates from the idea that even if we cannot observe simultaneously the outcome on the same unit receiving and not receiving the treatment, we can still define those potential quantities and build methods that allow their estimation.

Let us consider an outcome variable Y and a treatment T. We denote $T = 0$ and $T = 1$, respectively, the treated and the not treated condition. Then we can define $Y_i(T = 1)$ the potential outcome we would have observed if unit i had received the treatment and as $Y_i(T = 0)$ the potential outcome we would have observed if the same unit had not received the treatment. The methods that belong to the framework usually require the following assumptions:

Stable unit treatment value assumption (SUTVA). Applying the treatment to one unit does not affect the outcome of other units.

Unconfoundedness. The treatment assignment mechanism is conditionally independent of the potential outcomes given the covariates.

The most used methods developed in this framework include matching, instrumental variables, synthetic control and regression discontinuity designs.

Full matching

Here *full matching* (P. R. Rosenbaum 1991; Hansen 2004; Stuart and Green 2008) to estimate the treatment effect will be used. This technique groups all the units into a series of matched subclasses, containing at least one treated and one control unit. Similar units are gathered in the same subclass and its size depends on the number of comparable units: the more the available similar units, the larger the generated subclass and vice-versa. The similarity between units i and j is described by discrepancy measure δ_{ij}, which is usually calculated as a difference of distance measures, such as a propensity score. Small values of δ_{ij} indicate similar units, and thus as δ_{ij} increases the probability of i and j being matched decreases. If a pair $\{i, j\}$ has $\delta_{ij} = \infty$ the two units cannot be paired.

Let us consider a set of treated units $\mathcal{T}$, a set of control units $\mathcal{C}$ and a discrepancy measure $\delta_{ij} \in [0, \infty]$ computed for each pair $i \in \mathcal{T}$ and $j \in \mathcal{C}$. A full matching $\mathbf{S}$ maps the elements of $\mathcal{T} \cup \mathcal{C}$ into $\{0, ..., S\}$, where S is a postive integer that indicates the number of subclasses and $M = \mathcal{S}^{-1}[s]$, with $(1 \leq s \leq S)$, are the matched sets. Matched sets M are thus defined as the ensemble of units i and j assigned to a certain sunbclass s. The matching procedure is performed by minimizing the net discrepancy

$$\sum_{i \in \mathcal{T}, \mathcal{S}(i)>0} \sum_{j \in c, \mathcal{S}(i)=\mathcal{S}(j)} \delta_{ij}, \tag{2.3}$$

subject to
(i) $min(\#(M \cap \mathcal{T}), \#(M \cap \mathcal{C})) = 1 \quad$ and
(ii) $\forall i \in M \cap \mathcal{T}$ and $j \in M \cap \mathcal{C}, \quad \delta_{ij} < \infty$,
where $\#(M \cap \mathcal{T})$ and $\#(M \cap \mathcal{C})$ indicate respectively the number of treated and control units in M. The quantity in (2.3) represents the sum of discrepancies within each matched set M, summed over every matched set. Minimizing this sum reflects the idea of generating subclasses that gather similar units. The first constraint forces each matched set M to have at least one control and one trated unit whereas the second constraint imposes that unit pairs for which $\delta_{ij} = \infty$ cannot be placed in the same set.

Full matching has been chosen because it allows the estimation of the ATE, and thus the results of the matching procedure are coherent with the interventional do-notation defined in the context of causal graphs. Other matching procedures, in fact, only allow estimation of treatment effect on target subpopulation, such as the average treatment effect on the treated (ATT) or the average treatment effect on the controls (ATC). The do-notation instead focuses on comparing the average outcome resulting from applying the treatment to all the units and the average outcome we would observe if all the units were controls.

2.4 Analysis and results

The first stage of the analysis will consist in estimating a causal graph on the dataset to study the interactions between the considered variables. In a second step, given the structure of the obtained graph, a sufficient adjustment set will be selected via the back-door criterion. Lastly, the selected set will be used to implement a full matching procedure and estimate causal effects.

2.4.1 Learning the causal graph

Learning a causal graph from a dataset implies choosing a structural learning algorithm that will encode the structure of conditional independence of the data into a DAG.

A score-based algorithm called *Tabu Search* (Stuart Russell and Norvig 2002) has been implemented. The Tabu Search requires both faithfulness and causal sufficiency assumptions. As mentioned in Section 2.2, the graph will be learned on the dataset without employing sample weights. This choice originates from the fact that integrating weights in structural learning algorithms implies many methodological obstacles that are yet to be completely overcome. Solutions to use a weighting scheme in the learning procedure are still limited and currently being explored in the related literature. For some recent advances on the topic see Marella and Vicard (2022).

This part of the analysis has been carried out employing the *Bnlearn* package (Scutari 2010) in R Statistical Software (R. Core Team 2013). The procedure begins with computing the *BIC score* of an initial graph G, which is usually empty. The graph is then modified by adding, deleting or reversing its edges and a score is computed again for each of this variations. Note that this step is performed under the constraint that each modification cannot generate cycles. At the end of this phase Tabu Search retains the best scoring structure and use it as the new initial graph G. After performing each iteration the algorithm keeps track of the already-explored structures in a tabu list, so as not to compute the score of a given DAG twice. The procedure is repeated until a new iteration does not generate an increase in the best obtained score. When this happens the algorithm stops and the best scoring graph of the last iteration is selected as the algorithm output. Tabu Search has been selected because it is faster and more accurate than most algorithms for both small and large sample sizes (Scutari, Graafland, and Gutiérrez 2019).

Most structural learning algorithms, including Tabu Search, allow the inclusion of prior knowledge in the learning procedure by introducing constraints on the graph's structure. A known causal relationship between variables or the absence of it is encoded in the graph by imposing or forbidding a directed edge between two nodes. Tabu Search thus behaves as a supervised learning procedure where the graph is not only the result of the information emerging from the data but also of subject matter knowledge introduced in the form of constraints.

Prior knowledge of the subject matter has been synthesized in Table 2.2. The variables have been divided into four logical groups according to their type, as described in Section 2.2. The first group is *Precovid demographics* and contains primary firms' features such as their size, the geographical area where they operate and their business sector. The second group contains additional firms' traits that characterized them before the outbreak. The variation in their revenues and number of employees relative to the last three years, their past strategical activities and exposure to credit rationing are, among others, included in this category. The third group contains the variable which describes the firms' future expectations concerning revenues, measured prior to the pandemic. Post-covid features, such as the number of confirmed covid infections in their province and if they have been targeted by lockdown measures or not, belong to the fourth and last group.

The idea behind this logical categorization is that we assume that variables belonging to a specific group cannot affect the preceding groups described in Table 2.2. For example, a variable of the last group cannot cause variables of the other three groups, the third group cannot affect the first two and so forth. The four categories are then translated into constraints in the graph structure and used in the structural learning procedure.

Table 2.2. Logical variable groups

Precovid demographics ↚	Other precovid features ↚	Precovid expectations ↚	Postcovid features
Size (n. of employees)	Innovation, R&D	Pre-covid delta expcted revenues	Confirmed covid infections
Geographical area	Credit rationing		Essential business sector
Business sector	Export		
Manager education	Delta number of employees		
	Digital litteracy		
	Past delta revenues		

In addition to the mentioned constraints, some additional assumptions have been included in the model. Firstly, the treatment and the outcome variable are not allowed to cause any of the pre-treatment variables. This constraint originates from the fact that pre-treatment variable cannot be affected by the treatment or the outcome since they are measured before the treatment is applied. It is also assumed that *geographical area* cannot be affected by the other variables in the first group. Note that while it is assumed that variables of a group cannot cause variables of previous groups, according to the ordering contained in Table 2.2, variables belonging to the same group can affect each other. Lastly, we assumed that the variable *Essential business sector* could only be affected by the firms' business sector by definition.

The graph learnt by the algorithm is shown in Figure 2.3. Dashed arrows denote edges that have been forced to be present, based on assumptions regarding the existence of causal relations between variables. This set of assumptions based on prior knowledge, complements the forbidden arcs assumptions deriving from the logical categories in Table 2.2.

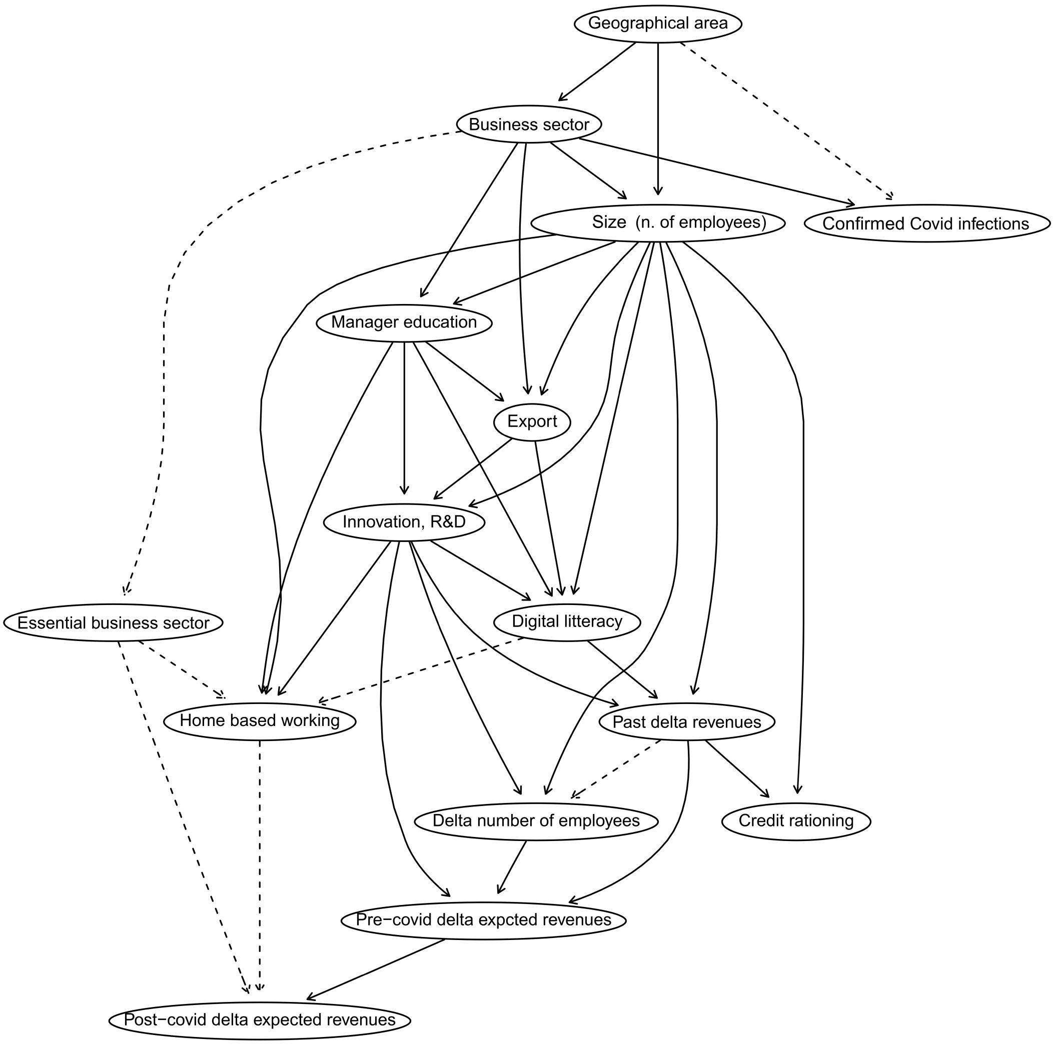

Figure 2.3. Causal graph learnt with the Tabu Search algorithm

The causal graph unveils the complex network of causal relations between the variables. The graph is dense, and the emerging structure highlights how all the different considered dimensions contribute to shaping the outcome variable *Post-covid delta expected revenues.* Note that apart from the dashed edges, the causal connections in the graph have not been imposed and emerge from the data under the assumptions we specified previously.

Structural variables nodes such as business sector, size and geographical area are the roots of the causal structure, consistently with the logical categorization imposed in the learning phase. This group shows causal connections among its elements and towards variables belonging to other groups. In particular, the graph reveals numerous directed edges between basic firm characteristics, strategies and performance, as one would expect according to economic intuition. For example, *geographical area* affects *business sector*, which in turn influences *size (n. of employees).* Certain regions are, in fact, more favourable to a specific type of economic activity, and the latter influences firm's size.

Geographical area and *confirmed covid infections* are related by construction since infections are measured at the province level. An arc between the two nodes has thus been imposed. The same is valid for *business sector* and *essential business sector. Business sector* also affects *manager education* and *export*, indicating how different economic activities require and are thus characterized by different levels of internationalization and education at the managerial level. The last variable affected by *business sector* is *confirmed covid infections.* For example, this relation could confirm that in certain territories, where a specific type of economic field is prevalent, characteristics related to that particular activity could have affected the spread of the disease.

The number of confirmed covid infections at the province level seems to not affect other variables in the model. This could be because, at the beginning of the pandemic, the perception of the covid related risk was shaped by covid infection numbers at the national level. Media outlets reported governmental declarations that focused on the evolution of the infection in the whole country. This kind of communication probably shifted the attention from infections at a local scale to infections at a national scale. Secondly the variable represents the number of detected covid infections and therefore constitutes a biased proxy of the real amount of infected individuals. The detected number of infections in fact under-measures the real infections and could potentially provide an untrue picture of the geographical heterogeneity in the spread of the virus.

The firm size is the structural feature that causes the largest number of variables in the model. The affected nodes include implemented strategies, revenues variation, employment variation, credit rationing and the implementation of home-based working. *Manager education* affects the implementation of several strategic choices such as creating innovative products or services, investments in research and development, export, digital literacy, and the implementation of HBW. In general, all the nodes that describe implemented strategies show dense connections among them, indicating that adopting or not adopting one of them also affects decisions concerning the implementation of the others.

Digital literacy of the firm, which depends on implemented strategies and structural features, is assumed to affect HBW. Giving employees the chance to work remotely requires proper training and infrastructure. An edge between essential business sector and treatment has also been imposed. This constraint follows the idea that the implementation of HBW must be affected by lock-down policies. Firms that were forced to shut down temporarily had a higher urgency to enable their employees to continue working from their homes. Past performance, captured by *past* Δ *revenues* and Δ *number of employees*, is directly caused by size of the firm and implementation of innovations and R&D. The revenues variation is also affected by *digital literacy.* The two performance variables have been assumed to be linked, and thus, an edge between them has been imposed. Note that even if only a few variables affect performance indicators directly, indirect connections show that many more dimensions take part in shaping their value. Almost at the end of the causal chain, we find pre-covid expected revenue variation node, directly caused by the vertices representing past revenues, variation in the number of employees, *innovation and R&D.* As expected, past performance is a strong driver of the beliefs concerning future performance. The outcome variable, *Post-covid delta expected revenues*, is caused by the treatment by assumption. Moreover, it has been assumed that the outcome is affected by the variable *essential business sector*, which denotes if a firm was targeted or not by lock-down policies. This imposed relation translates the idea that being forced to close temporarily has negative consequences on future revenues most of the time. Lastly, the graph shows that expectations after the spread of the virus are affected by pre-covid expectations. This relation highlights that even if the pandemic had a strong negative impact across all businesses, the reaction to this shock depends on the firm conditions before the covid outbreak.

2.4.2 Adjustment set selection and estimation of the ATE

The obtained causal graph is used to select a sufficient adjustment set of covariates to estimate the effect of implementing home-based working on expected revenues. The R Statistical Software (R. Core Team 2013) package *Dagitty* (Textor et al. 2016) has been employed. Given the graph in Figure 2.3, the minimal set which satisfies the back-door criterion for the effect of T on Y is

$$S_{adj} = \{\textit{Essential business sector}; \textit{Pre-covid delta expected revenues}\}$$

The two variables block all the confounding paths between treatment and outcome and thus allow the estimation of unbiased treatment effects.

Full matching is employed for causal effect estimation. The methodology has been implemented in R Statistical Software (R. Core Team 2013), using the package *MatchIt* (Stuart, King, et al. 2011). The first phase of the procedure involves assessing the balance of covariates after matching. The results of the procedure are shown in Figure 2.4. Covariate balance improves considerably after matching. Once propensity score is estimated and balance is achieved, the average causal effect of T on Y is calculated by regressing the outcome on the treatment and the adjustment set in a weighted regression model, also referred to as the *outcome model.* The fitted model is then used to predict the distribution of the outcome if all units were controls and if all units were treated. This kind of procedure, also called *g-computation* (Snowden, Rose, and Mortimer 2011), is required when we include additional covariates in the outcome model and we are interested in estimating a marginal effect. The obtained distribution of the outcome under treatment administration Y_1 and control administration Y_0 are then averaged and used to compute the causal risk ratio

$$ATE = \frac{E[Y_1]}{E[Y_0]}. \tag{2.4}$$

Note that denoting the interventional distribution of Y with Y_x is typical of the potential outcome framework (Imbens and Rubin 2015) and is equivalent to Pearl's do-notation $P(Y|do(X = x))$. Standard errors are computed through block bootstrap (Abadie and Spiess 2020).

The estimated ATE in Table 2.3 shows that implementing HBW from the beginning of the pandemic helps mitigate the harmful effects of Covid-19. In particular, treated firms have a higher probability of expecting stable or increasing future revenues and a lower probability of a strong decrease. For every point estimate, confidence intervals (C.I.) have been estimated with bootstrap at a 95% confidence level. The different width of confidence intervals is primarily due to an uneven frequency distribution in the outcome variable levels. The estimated causal risk ratio for increasing and stable expected revenues, with 95% C.I. between parenthesis, is respectively 1.8 (0.96, 3.13) and 2.2 (1.67, 3.03). The probability of a decrease between -15% and 5% in expected revenues is almost the same for treated and controls, with an ATE of 0.97 (0.61, 1.43). In contrast, the ratio equals 0.73 (0.56, 1.07) for an expected decrease lower than -15%.

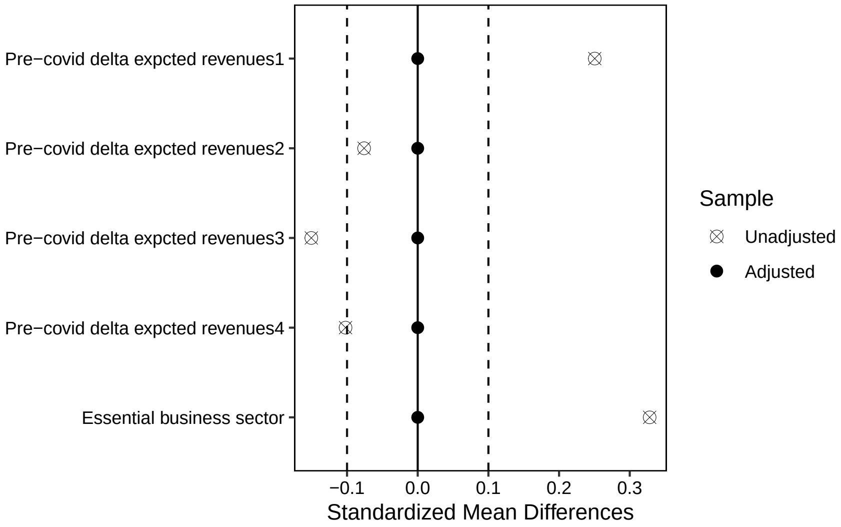

Figure 2.4. Balance of selected covariates

Table 2.3. ATE estimates of HBW implementation on post-covid Δ expected revenues

Post-covid delta expected revenues	Point Estimate	95% C.I.
Increase (>+5%)	1.80	(0.96, 3.13)
Stable (between -5% and +5%)	2.11	(1.67, 3.03)
Decrease (between -15% and -5%)	0.97	(0.61, 1.43)
Strong decrease (<-15%)	0.73	(0.56, 1.07)

2.5 Discussion

The poposed combined methodology allows the construction of a causal graph, the selection of a sufficient set of variables for adjustment, and the ATE estimation. Causal relations emerging from the graph are coherent with economic intuition and reveal new insights concerning variables interactions. The graph leads to selecting an adjustment set composed of *business sector* and *pre-covid* Δ *expected revenues*. According to the graph's structure, the two variables block all the spurious paths between treatment and outcome, thus ensuring unbiased ATE estimation. The resulting causal effect, estimated with full matching, shows a positive impact of HBW on expected future revenues. Receiving the treatment, ceteris paribus, corresponds to having a higher probability of stable or increasing expected revenues.

The first contribution of this work to the literature is the proposed methodology. Causal graphs are yet to be the most used approach in economic causal inference literature, but we believe they constitute an irreplaceable resource. Learning a causal graph from data is a process that translates causal information into a more transparent medium. The graph constitutes itself a set of assumptions concerning the causal relationships between variables on which causal estimates are based. Encoding information into a causal diagram thus improves the analysis's clarity and understandability of how results are derived. In addition, the graph learning step allows the introduction of prior knowledge into the model, imposing theory or evidence-based relations between variables. Estimation of ATE using the adjustment set selected from the graph via back-door criterion can be then performed with a method of choice, such as simple regression or matching, depending on the assumptions being made. Regardless of the chosen method, using the graph-selected adjustment set ensures unbiased ATE estimation if the implied assumptions are satisfied.

The second contribution of this work resides in the results of the analysis. The treatment causes a change in the distribution of the outcome that partly counterbalances the impact of the Covid-19 shock. In other words, the outcome variation generated by the treatment always goes in the opposite direction of the observed change induced by the pandemic outbreak. This finding is coherent with HBW literature and, in particular, with its impact on flexibility and productivity. However, this mitigating effect was yet to be quantified in a comprehensive causal framework.

Future research could focus on measuring the effect of HBW on different performance indicators and after different periods of time. Moreover, additional analyses will be carried out in order to check if the findings hold for a wider population of firms. Assessing causal effects on overall performance was beyond the scope of the analysis, but the obtained results contribute to having a better understanding of the implications of working from home in Covid times. Reconstructing the complete picture is needed to guide future policies and assess the urgency of providing the firms with the necessary resources for HBW implementation.

Chapter 3

Bootstrap-aggregated adjustment set selection

Abstract

Causal effects can be estimated from observational data within the Causal Graph framework. When the true causal graph is unknown, data are used to learn the graph through structural learning algorithms. The obtained model is then employed to select a sufficient set of covariates for adjustment, according to the back-door criterion. Graph learning is a crucial step of the process since misspecification in the graphical model can lead to incorrect adjustment set selection and thus generate biased estimates. We propose a procedure that resorts to bootstrap-aggregating to select the adjustment set. First, an ensemble of graphs is learnt on bootstrapped replicates of the original dataset, and then a multiset of adjustment sets is obtained by applying the back-door criterion to each graph of the ensemble. Finally, the element with the highest multiplicity in the multiset is selected as the resulting adjustment set. The simulations on graph structures of low complexity reveal that, at small sample sizes, the novel procedure is less accurate than the benchmark methods. However, as the graph complexity increases, the relative performance of the proposed method improves. When applied to complex graphical structures, bootstrap-aggregated adjustment shows the highest accuracy among the tested methods for both small and large sample sizes.

3.1 Introduction

Causal inference investigates causal relations between variables. Questions like "what happens to A if I intervene on B?" are causal and must be answered through models that follow the rules of causality. The gold standard of causal inference is experimental data since it allows a direct comparison between treated and control units. Thus an estimate of the causal effect of the treatment can be obtained without further calculation. Unfortunately, conducting randomized experiments is usually too expensive or impossible. This is often the case in epidemiology or economics, where data is primarily observational. The need of dealing with observational data has led to many methods capable of estimating causal effects even when experimental data is not available.

One of the possible frameworks to investigate causal claims, when dealing with observational data, are Causal Graphs (Pearl 1995), also called Causal Diagrams or Causal Bayesian Networks. In this approach, the causal model is composed of a graph, which entails the relations between variables in the form of nodes and edges and a joint probability distribution. Causal diagrams can handle models with many variables, they represent variable interactions clearly through a graph, and, if identifiable, estimate the treatment's causal effect on an outcome variable. In order to compute the causal estimate starting from a known causal diagram, one can resort to the rules of do-calculus, as explained in Pearl, M. Glymour, and Jewell (2016).

An alternative and widely used approach is the Potential Outcomes framework (Splawa-Neyman, Dabrowska, and Speed 1990; Rubin 2005). The methods belonging to this approach are useful to compute causal effects from experimental data and, under specific circumstances from observational studies, resorting to particular identification strategies such as matching, instrumental variables, synthetic control methods and regression discontinuity designs. The reasoning behind these techniques aims at checking if the main features of randomized experiments still hold or can be emulated in some special cases of observational studies (Imbens 2020).

When the treatment assignment mechanism is unknown, the main concern for both approaches is confounding: the situation where there is at least one variable that has a direct causal effect on both the treatment and the outcome. This kind of configuration can generate biased estimates unless adequately accounted for. However, if the causal relations between variables are known a priori, it is possible to select a sufficient set of covariates for confounding adjustment, if such a set exists. In the Causal Graphs framework, sufficient adjustment sets can be derived from the structure of the graph, following the back-door criterion (Pearl 1995). Once the set is selected, adjustment can be performed through various methods, including matching, weighting, regression adjustment or doubly robust methods (Abadie and Cattaneo 2018).

Unfortunately, the true graph is often unknown in real settings or just partially known and must be estimated from data. Several algorithms have been developed to solve the task, and a review of their performance can be found in Scutari, Graafland, and Gutiérrez (2019) and Constantinou (2020). Once the graph is estimated, it is usually taken as a good approximation of the true graph and then a sufficient adjustment set is selected to estimate a chosen treatment effect. The technique, however, depends on the estimated graph, and if the structure of the graph is misspecified, the selected adjustment set could be not sufficient for confounding adjustment.

This work proposes a procedure to select the adjustment set from multiple graphs, employing bootstrap. In particular, the method consists in generating several bootstrap replicates of the original dataset, then on each of them learn a graph and select an adjustment set via back-door criterion. Finally, the obtained sets are put together to form a multiset, and the set with the highest multiplicity is chosen for adjustment. Generating multiple versions of a predictor by bootstrapping and then building an aggregated predictor is also called bootstrap-aggregating or bagging (Breiman 1996). Bagging has been applied to classification and regression trees, subset selection in linear regression (Breiman 1996), non-parametric regression (Borra and Di Ciaccio 2002), clustering (Dudoit and Fridlyand 2003) and neural networks (Ha, Cho, and MacLachlan 2005). However, to the author's knowledge, this approach has never been used to select adjustment sets from causal graphs.

The proposed procedure is tested on three graphs of increasing complexity and then compared to other adjustment set selection techniques. All the procedures show high accuracy levels when tested on the simplest graph. Bootstrap-aggregating the adjustment set makes no exemption, even if its accuracy is lower than the benchmark methods at low sample sizes. As complexity increases, the overall accuracy of the procedures decreases, but the relative performance of bagging improves. In the simulations on the medium complexity graph, the novel procedure achieves the same level of accuracy as the other graphs, whereas on the most complex graph, bagging has the highest accuracy among the chosen methods, at both small and large sample sizes.

The paper is organized as follows. Section 3.2 introduces some basic causal graph terminology and notation, which is necessary for explaining how the procedure works. The proposed method is described in detail in Section 3.3, while Section 3.4 contains simulations and results. Findings and room for future work are discussed in Section 3.5.

3.2 Background and notation

3.2.1 Causal graphs

A graph $G = (V, E)$ is a collection of nodes or vertices V and edges E. A *causal graph* is a graph where nodes represent random variables and edges describe the causal relations between these nodes. If two nodes X_i and X_j are connected by an edge the nodes are *adjacent*. A *path* between two nodes X_i and X_j is a sequence of nodes beginning with X_i and ending with X_j where all the nodes are connected to the next. Edges can be *directed* or *undirected*. An edge is directed if it goes out from one node into another and undirected without such orientation. A graph where all the edges are directed is a *directed graph*. If a directed edge goes from X_i to X_j then X_i is a *parent* node of X_j, and X_j is a *child* of X_i. We will denote parents of node X_j with the notation pa_{X_j}. A *directed path* between two nodes is a path where no node has two edges on the path directed into it, or two edges directed out of it. Given a directed path, the first node is an *ancestor* of every node of the path, and every node of the path is a *descendant* of the first node. A directed path that begins and ends with the same node is a *cycle.* A directed graph containing no cycles is called a *directed acyclic graph* (DAG). When two nodes X_i and X_j point to a third node X_k and X_i and X_j are not connected by an edge, X_k is a *collider* in the ordered triplet of nodes (X_i, X_k, X_j). The ordered triplet with a collider as the middle node is also called *unshielded triplet.*

A DAG encodes statements of conditional independence through the notion of *d-separation* (Pearl 2000, Definition 1.2.3, page 16). Consider a DAG G with nodes $\mathbf{X}$, two nodes X_i and X_j belonging to $\mathbf{X}$ with $X_i \neq X_j$ and a set of nodes $\mathbf{S} \subset \mathbf{X}$ not containing X_i and X_j. Then X_i and X_j are d-separated given $\mathbf{S}$ in G, if there is no path p connecting X_i and X_j such that (i) every collider on p has a descendent in $\mathbf{S}$ and (ii) no other node on p belongs to $\mathbf{S}$. If two nodes are d-separed by a set $\mathbf{S}$ then they are conditional independent given $\mathbf{S}$. If two vertices are instead d-separated without conditioning on a set, they are said to be marginally independent.

Given a causal graph G with ensemble of nodes $\mathbf{X}$, every $X \in \mathbf{X}$ is independent of all its non-descendants, conditional on its parents pa_X. This implies that the joint probability distribution of the nodes $P(\mathbf{X})$ can be factorized as follows

$$P(\mathbf{X}) = \prod_{X \in \mathbf{X}} P(X|pa_X) \tag{3.1}$$

If a distribution allows the factorization in 3.1, according to the structure of the graph, the graph is said to satisfy the *Causal Markov condition* with respect to the distribution.

Note also that each node is always conditionally independent of every other node of the network given a set consisting of its parents, its children and the parents of its children. This set is also called the *Markov blanket* of a node (Pearl 2009b) and its average size constitutes a measure of complexity of the conditional independence structure of the graph.

3.2.2 Graph learning

If the underlying dependence model of a problem is unknown, the graph can be estimated from a dataset containing the variables of interest (Koller and Friedman 2009). Most of the procedures that learn graphs from data require the following assumptions (Spirtes et al. 2000):

Faithfulness. A probability distribution P is faithful to a causal graph G only if the set of conditional independences of P is exactly the same as those described by G. In the context of graph learning, this implies having an exact correspondence between the conditional independence relations of the distribution of the data from which the graph is learnt and those encoded in G through d-separation rules. The assumption also ensures that the distribution describes no extraneous independences with respect to those entailed by the causal Markov condition applied to G.

Causal sufficiency. There is no unobserved variable in the model that would cause the causal Markov condition to be violated.

When the required conditions are satisfied, a causal graph can be retrieved from data, through a *structural learning algorithm.* The task of recovering a graph G from a dataset D containing N observations through a learning algorithm can follow one of three possible approaches: *constraint-based*, *score-based* and *hybrid.* Constraint-based algorithms perform a sequence of conditional independence tests. The resulting conditional independence structure is then translated into graphical form through d-separation rules and encoded in a DAG G. *Score-based* algorithms instead assign to each candidate DAG a score reflecting its goodness of fit and select the network which maximizes it. *Hybrid algorithms* combine the two approaches using conditional independence tests to restrict the space of possible networks structures and then select the DAG that maximizes a given network score within the restricted space.

3.2.3 Calculus of intervention

This work focuses on the estimation of the effect of an intervention from observational data. The transition from an observational context to an interventional one requires some further discussion. Pearl (2009a) introduces the notation $do(X_i = x_i)$ to indicate that the variable X_i is set to the value x_i by intervention. In this way we can denote as $P(X_j|do(X_i = x_i))$ the distribution of X_j given that X_i is forced to take value x_i. If we assume that X_i is a discrete random variable which can take value 0 or 1, the average causal effect (ACE) of an intervention on X_i can then be denoted as follows (Imbens 2004)

$$ACE = E(X_j|do(X_i = 1)) - E(X_j|do(X_i = 0)).$$

In an observational context the quantity $P(X_j|do(X_i = x_i))$ is not measured but it can be derived if the following assumptions are satisfied (Pearl 2000). Given a graph G, a treatment variable T, an outcome variable Y and a set of covariates $\mathbf{S} \subseteq \mathbf{X}$, where $\mathbf{X}$ denotes the ensemble of the nodes of G,

Assumption 1. *Every $s \in \boldsymbol{S}$ is a nondescendant of T*

Assumption 2. *All back-door paths from T to Y are blocked by $\boldsymbol{S}$,*

A set $\mathbf{S}$ which satisfies Assumption 1 and 2 is called a *sufficient* adjustment set (Greenland, Pearl, and Robins 1999).

Back-door paths, mentioned in Assumption 2, are defined as paths between T and Y that contain an arrow into T. Consider the DAG in Figure 3.1. The graph shows the presence of a back-door path p between the pair (T, Y), traced along the ordered tuple $\{T, X_1, X_2, Y\}$. These are spurious paths, and if they are not blocked by controlling for a sufficient adjustment, causal effect estimation could produce biased results.

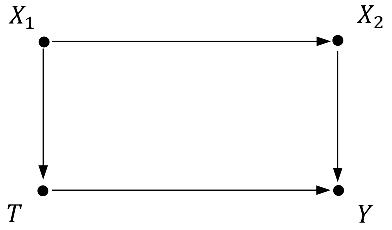

Figure 3.1. A DAG with a back-door path

Assumption 1 and 2 together form the *back-door criterion.* Checking if a set satisfies the criterion can be tested by systematic graphical procedures that are applicable on graphs of any size and complexity (Pearl 2000). A generalized version of the back-door criterion also exists and it permits to check if a set is sufficient for adjustment on several classes of graphical models (Perković et al. 2015). Note that the definition of $\mathbf{S}$ allows the set to be non-unique. Several adjustment sets could in fact satisfy Assumptions 1 and 2, thus being all sufficient for causal effect estimation.

If a sufficient adjustment set exists, then interventional distribution can be expressed in observational terms:

$$\begin{aligned}
&P(X_j|do(X_i = x_i)) \\
&= \sum_{S} P(X_j|\mathbf{S} = \mathbf{s}; do(X_i = x_i))P(\mathbf{S} = \mathbf{s}; do(X_i = x_i)) \\
&= \sum_{S} P(X_j|\mathbf{S} = \mathbf{s}; X_i = x_i)P(\mathbf{S} = \mathbf{s})
\end{aligned}$$

The absence of the do-notation in the last line implies that causal effects can be estimated from observational data when the assumptions hold. This is a central result for causal estimation through graphical models, and it will be a fundamental building block of the proposed procedure.

3.3 Bootstrap-aggregated adjustment set

Consider a dataset $D = (\mathbf{X}; T; Y)$ of size n sampled from an unknown graph G containing a set of covariates $\mathbf{X}$, an outcome Y and a treatment T. Since we are interested in computing the effect of T on Y, we need to select an appropriate sufficient adjustment set to remove confounding bias. A straightforward way to proceed would be estimating a graph $\hat{G}$ through a structural learning algorithm $\mathcal{L}$ on D and then selecting a sufficient adjustment set according to the generalized back-door criterion. This procedure, however, relies entirely on the precision of $\mathcal{L}$ to retrieve the correct network structure from D. In particular, if some of the local relations between T and Y are misspecified in $\hat{G}$, the adjustment set resulting from applying the back-door criterion on $\hat{G}$ could not be sufficient for G and thus produce biased causal estimates.

We propose a novel method that investigates how adjustment set selection would vary if we modify D through bootstrap (Tibshirani and Efron 1993) and then selects a single adjustment set according to its stability.

3.3.1 Description of the method

Bootstrap is a resampling technique that creates replicates of the original dataset by sampling from it with replacement. The procedure is repeated M times to produce M bootstrap replicates of the same size as the original dataset. This work builds on a particular implementation of bootstrap called bootstrap-aggregating or bagging (Breiman 1996). Bagging is a machine learning method that generates multiple versions of a predictor through bootstrap and uses these to get an aggregated predictor. A plurality vote is performed if the predicted element is a class. Here, the logic of bagging is applied to graph structural learning and adjustment set selection. Bootstrap samples are used to fit M different models and derive a multiset Θ of adjustment sets. Then the element of the multiset with the greatest multiplicity $\theta^\star$ is selected. Note that given an estimated graph $\hat{G}$, more than one set of variables can satisfy the generalized back-door criterion, and therefore the elements of the multiset can be greater than the number of bootstrap replicates. The pseudocode of the procedure is contained in Algorithm 3.

In the first step, a sample D of size n is used to produce M bootstrap replicates of the same size. The bootstrap procedure creates new datasets by sampling with replacement from the original sample until reaching the target sample size n. Therefore, all the units contained in the obtained replicates $B_1, ..., B_M$ appear in D.

In step 2, the multiset Θ is initialized as an empty set, which will be populated in the following iterations. A graph $\hat{G}_i$ is learnt with an algorithm $\mathcal{L}$ on each generated bootstrap sample B_i, given the constraint that the treatment T is a parent of the outcome Y. This assumption encodes in the graph the a priori knowledge that T directly causes Y. Note that $\mathcal{L}$ is defined as a generic learning algorithm because the method supports the implementation of any learning procedure that generates a

Algorithm 3: Bootstrap-aggregated adjustment set

```
Input: A sample $D = (\mathbf{X}, T, Y)$ from an unknown graphical model $G$, a
       number of bootstrap samples $M$, a structural learning algorithm $\mathcal{L}$,
       an adjustment set selection procedure $\mathcal{A}$.
Output: A bootstrapped adjustment set $\theta^\star$
1 Generate $M$ bootstrap samples $B_1, ..., B_M$ from $D$;
2 $\Theta = \emptyset$;
3 for $i = 1$ to $M$ do
4     $\hat{G}_i = \mathcal{L}(B_i | T \in pa_Y)$;
5     add $\mathcal{A}(\hat{G}_i, T, Y)$ to $\Theta$;
6 end
7 $\theta^\star = \max_{\theta_j \in \Theta} \{v(\theta_j)\}$
```

DAG.

Then an adjustment set selection method $\mathcal{A}$ for estimating the effect of T on Y is applied to each graph $\hat{G}_i$. The selected sets are then added to Θ until steps 4 and 5 have been iterated over all the bootstrap resamples. Also, the adjustment set selection criterion $\mathcal{A}$ is not specified because the procedure allows the implementation of different methods. Since given a graph, a treatment and an outcome variable, more than one adjustment set can satisfy the back-door criterion, several studies have been carried out to assess how to select one set or another. One of the most common method targets sets with the minimal cardinality among all admissible sets, while others focus on selecting the set with the smallest asymptotic variance (Witte et al. 2020).

The last step of the algorithm defines the bagged adjustment set $\theta^\star$ as the set which maximizes $v(\theta_j)$ among all sets $\theta_j \in \Theta$, where $v(\cdot)$ denotes the multiplicity of a given element of a multiset. The bagged adjustment set $\theta^\star$ is thus the set, among all $\theta \in \Theta$, which satisfies the back-door criterion in the highest number of learnt graphs. Note that adjustments sets are selected according to the structure of the graph and in particular according to the confounding paths between Y and T. If a graph is large enough, changing some parts of it could leave the local configuration that define confounding paths between a given T and Y intact, thus not affecting the selected adjustment set. It follows that different graphs, learnt on different bootstrap replicates, could however generate the same adjustment set, because they share the same local configurations around treatment and outcome. Moreover note that even if two graphs show different confounding paths for a given pair (T, Y), there could however exist one or more adjustment sets that are sufficient for confounding adjustment in both graphs.

3.4 Simulation

In this section, we assess the performance of the proposed method by comparing it to three alternative approaches. The different techniques will be tested on graphs of increasing complexity, and the accuracy of the obtained results will be compared. The simulation has been carried out employing the *Bnlearn* (Scutari 2010) and the *Dagitty* (Textor et al. 2016) packages in R Statistical Software (R. Core Team 2013).

We will first describe how the accuracy measure is computed and how the simulation will be carried out. We will then provide details concerning the alternative benchmark procedures, and finally, the obtained results will be presented.

3.4.1 Simulation details

Given a known graph $G(\mathbf{X}, T, Y)$, where $T \in pa_Y$, we sample K datasets $D_1, ..., D_K$ of size n from it. Then we apply the proposed method on each sample D to obtain K adjustment sets $\theta_i^\star$ for the estimation of the causal effect of T on Y. Finally, we check if $\theta_i^\star$ is a sufficient adjustment set for calculating the effect of T on Y in the true graph G. The results are summarized by the quantity

$$R = \frac{\sum_{i=1}^{K} I_G(\theta_i^\star)}{K}, \tag{3.2}$$

where $I_G(\theta_i^\star)$ is an indicator function which takes value 1 if $\theta_i^\star$ is a sufficient adjustment set for (T, Y) in G and 0 otherwise. The formula of Equation 3.2 represents a measure of the precision of the proposed method since it computes the proportion of samples for which the obtained results are correct. Note that once the samples $D_1, ..., D_K$ are obtained from G, the true graph is considered unknown for the whole simulation procedure, as if the sampled data were the only available source of information. At the end of the simulations, G is employed again to build the accuracy measure R and evaluate the obtained results.

R is calculated for each pair (T, Y), such that $T \in pa_Y$ according to the structure of G. The total number of pairs (T, Y) is equal to $\sum\limits_{v \in \mathcal{V}} \#pa_v$ where $\mathcal{V}$ is the ensemble of vertices of graph G and $\#pa_v$ is the number of parents of vertex $v \in \mathcal{V}$. Once R has been calculated for every possible (T, Y) pair, an average of the results is computed to obtain a summary of the accuracy measure for the whole graph. If we denote $M = \sum\limits_{v \in \mathcal{V}} \#pa_v$, we can write the average of the accuracy measures for a given graph as

$$\overline{R} = \frac{\sum_{m \in M} R_m}{M}. \tag{3.3}$$

Equation 3.3 thus gives an account of the performance of an adjustment set selection procedure over the ensemble of nodes of a causal graph. In particular, it describes how well the technique succeeded in recovering a sufficient adjustment set from the data for every possible pair of treatment and outcome according to the graph structure.

The simulations are performed on three discrete networks of increasing size and complexity that have been frequently used in the literature: Asia (Lauritzen and Spiegelhalter 1988), Alarm (Beinlich et al. 1989) and Insurance (Binder et al. 1997). We set the number of samples $K = 10$, and the number of bootstrap replicates $M = 200$. Simulation with different values for the two parameters have been tested and these values have been chosen because they represent the best compromise between performance and computational time. The same number of replicates will be used in the bootstrap-aggregated procedure and in the benchmark methods that implement bootstrap. Different sample sizes are used in the simulations, according to the complexity of the network. The main characteristics of the chosen graphs, such as the number of nodes, edges, total parameters and average Markov blanket size, are described in Table 3.1.

Table 3.1. Characteristics of the graph employed for the simulation

	Nodes	Edges	Parameters	Average Markov blanket size
Asia	8	8	18	2.50
Alarm	37	46	509	3.51
Insurance	27	52	984	5.19

The structural learning algorithm $\mathcal{L}$ used to learn the graphs $\hat{G}$ is the *Tabu Search* algorithm (Glover 1986; Russell and Norvig 2009) with a BIC score. Tabu Search belongs to the family of score-based algorithms, and it has been chosen because it is more accurate and faster than most other learning algorithms, for both small and large sample sizes (Scutari, Graafland, and Gutiérrez 2019). The procedure's first step consists of computing the score of an initial, usually empty, graph. In the second step the score is computed again for every possible arc addition, deletion or reversal in the initial graph, which would still generate a DAG. The best scoring structure is retained together with its score and then the second step is repeated as long as the obtained best score increases. While computing the score of the possible configurations, the algorithm keeps track of previously-explored structures in a tabu list to avoid considering the same structure twice in different iterations. When an iteration fails to provide an increased score, the algorithm stops and the obtained DAG is selected as the output of the algorithm.

The chosen adjustment set selection procedure $\mathcal{A}$ is the *minimal adjustment set*. The method selects the globally minimal adjustment sets between all adjustment sets which satisfy the generalized back-door criterion. An adjustment set is globally minimal if it has the smallest cardinality between all possible adjustment sets. If more then one set share the same cardinality, which is also the smallest cardinality, then they all constitute a minimal adjustment set.

3.4.2 Benchmark methods

Bootstrap-aggregated adjustment performance is compared with three alternative methods. These methods focus on recovering a graph structure from data and then use the learnt diagram to select an adjustment set. The selected benchmark procedures have been identified by the literature as the most reliable in recovering the structure of the true graph from a dataset (Broom, Do, and Subramanian 2012).

Single algorithm graph

The first benchmark method is the most straightforward to apply and is often found in the literature. The procedure consists in learning a graph $\hat{G}_i$ from each sample D_i with a learning algorithm $\mathcal{L}$, thus obtaining K graphs $\hat{G_1}, ..., \hat{G_K}$. Then a sufficient adjustment set $\theta_i^\star$ for the effect of T on Y in each $\hat{G}_i$ is formed by selecting pa_T. Note that the treatment parents always constitute a valid set for causal effect adjustment. This method is less computationally intensive than bagging, but it relies entirely on the precision of $\mathcal{L}$ to recover the structure of the true graph G from the original samples $D_1, ..., D_K$.

Average graph

The second benchmark procedure involves bootstrapping to obtain an average graph and then selecting an adjustment set according to its structure. In the first step we sample M bootstrap replicates $B_{i1}, ..., B_{iM}$ from each sample D_i and then a graph $\hat{G}_{i,j}$ is learnt on each bootstrap replicate B_{ij} with learning algorithm $\mathcal{L}$. In the following step, the obtained graphs $\hat{G}_{i,j}$ are used to build a measure of confidence on all the arcs which appear in the graphs, based on how many times they appear. All the edges that show a confidence level higher than a certain threshold are included in a graph $\hat{G}_i^{avg}$ called the average graph. The procedure to build $\hat{G}_i^{avg}$ is explained in detail in Friedman, Goldszmidt, and Wyner (2013) and Imoto et al. (2002). Finally, given T and Y, an adjustment set $\theta_i^\star$ is composed by pa_T, according to the structure of $\hat{G}_i^{avg}$.

The difference between this method and the bootstrap-aggregated adjustment set is that the former does not integrate graph learning and adjustment set selection as the latter does. In the novel procedure we propose here, adjustment set selection is repeated on each bootstrap replicate, and then all the sets are combined, whereas, in the average graph method, bootstrap only concerns graph learning.

Best score graph

The third and last benchmark method also involves bootstrapping, but instead of generating an average graph, it selects the diagram with the best score among the bootstrap replicates. Initially M bootstrap replicates $B_{i1}, ..., B_{iM}$ are sampled from each sample D_i. Then we fit a graph $\hat{G}_{i,j}$ through $\mathcal{L}$ on each replicate B_{ij} and select the graph $\hat{G}_i^{bs}$ with the best score among all the bootstrap replicates for each sample D_i. Then an adjustment set is selected by setting $\theta_i^\star = pa_T$, according to $\hat{G}_i^{bs}$. The idea behind this method is that the graph with the highest score is the one that better describes the causal information contained in the data and could thus generate more reliable adjustment sets.

3.4.3 Results

The smallest graph employed for testing the procedure is the *Asia* network. The graph was first introduced by Lauritzen and Spiegelhalter (1988) and has now become a standard for testing structural learning algorithms. The diagram, represented in Figure 3.2 describes the interactions between lung diseases, symptoms and visits to Asia. However, recovering the model's structure is challenging for most algorithms, despite the apparent graph simplicity.

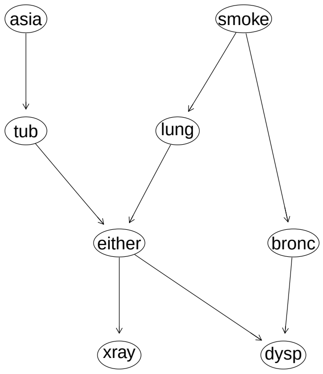

Figure 3.2. The *Asia* graph

The results are summarized in Figure 3.3. The x-axis represents sample size n, while the y-axis describes the average accuracy measure $\overline{R}$. All the procedures show good levels of accuracy, even at small sample sizes. At $n = 150$ the bootstrap-aggregated adjustment set records the lowest $\overline{R}$ (0.8), whereas the average graph procedure achieves the best performance (0.96). However, as sample size increases the performance of bagging improves and at $n = 1400$ all the methods achieve $\overline{R} = 1$, which corresponds to recovering a sufficient adjustment set for all the extracted samples and possible pairs of T and Y.

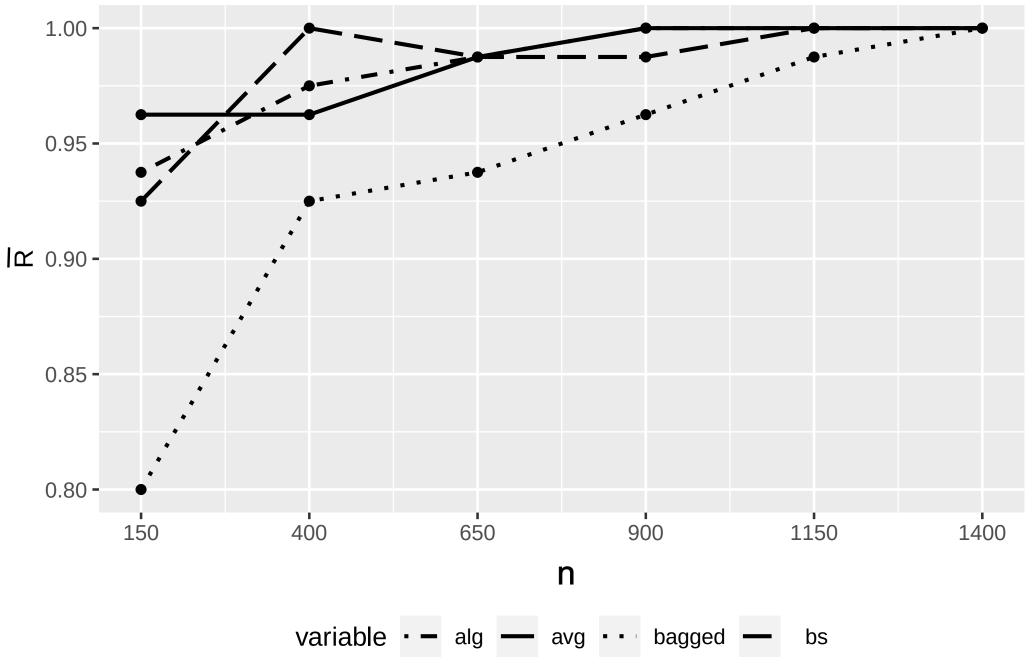

Figure 3.3. Results of the simulations on *Asia* network

The second graph on which the methods are tested is the *Alarm* network. This graphical model is one of the most studied in the structural learning literature and was first introduced in Beinlich et al. (1989). The network, represented in Figure 3.4, describes an intensive care patient monitoring system and consists of 37 discrete nodes, with two, three or four states.

The results for different sample sizes are shown in Figure 3.5. At the lowest sample size, $n = 200$, the procedure which has the highest accuracy is the average graph (0.81), followed by the bagged adjustment set (0.79) and the simple graph (0.78), whereas the best score graph is the least accurate (0.77). As n increases, computed $\overline{R}$ increases for all the procedures and becomes substantially stable for $n > 1400$. Even if all the methods show similar performances, bagging has the steepest accuracy increase and the overall best R as n grows larger.

The last and most complex test graph is the *Insurance* network. The model describes the risk evaluation mechanism of a car and was first found in Binder et al. (1997). The graphical model, shown in Figure 3.6, has 27 nodes, 52 edges and 984 parameters.

Selecting a sufficient adjustment set through the proposed procedures is way

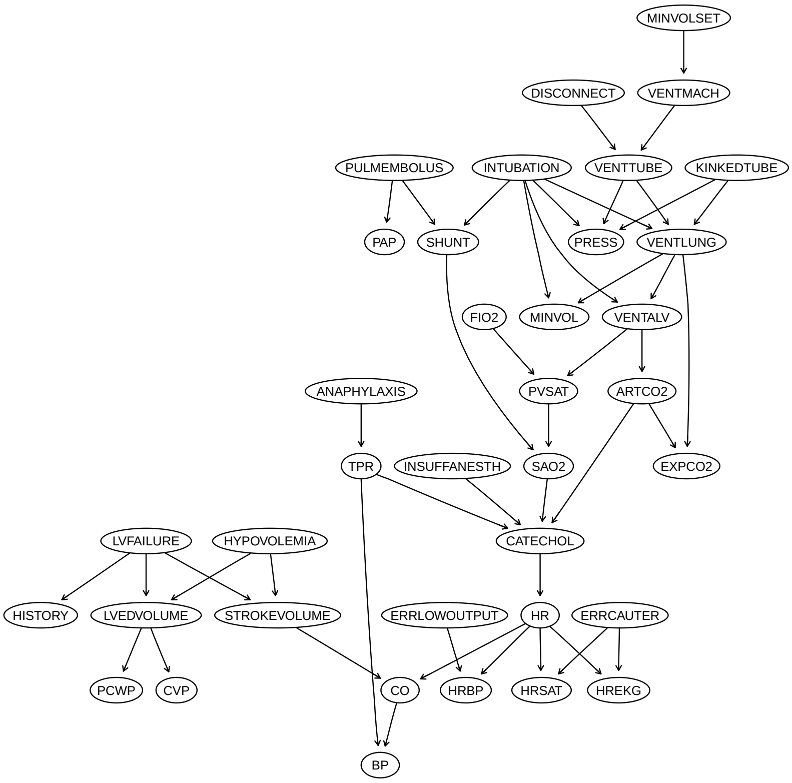

Figure 3.4. The *Alarm* graph

harder on a graph of this size, as the results in Figure 3.7 reveal. For n=200, the smallest sample size, $\overline{R}$ ranges between 0.12 and 0.23 and the bagged adjustment set has the highest accuracy $R = 0.23$. As n increases, the performance of the different procedures improve and bagging the adjustment set remains the most performing method. Among the three chosen graphs, the *Insurance* network is the one on which the novel procedure performs best, indicating a better accuracy as complexity increases relatively to the benchmark procedures.

Computational times are similar for all the procedures employing bootstrap resampling, whereas the single algorithm graph procedure is faster since it selects the adjustment set according to a single graph. However, among the bootstrap-based procedures, bootstrap-aggregating is slightly slower than the other ones. In general, computational times increase as the complexity of the true graph increases and, relatively to bootstrap-based procedures, as the number of bootstrap replicates grows larger.

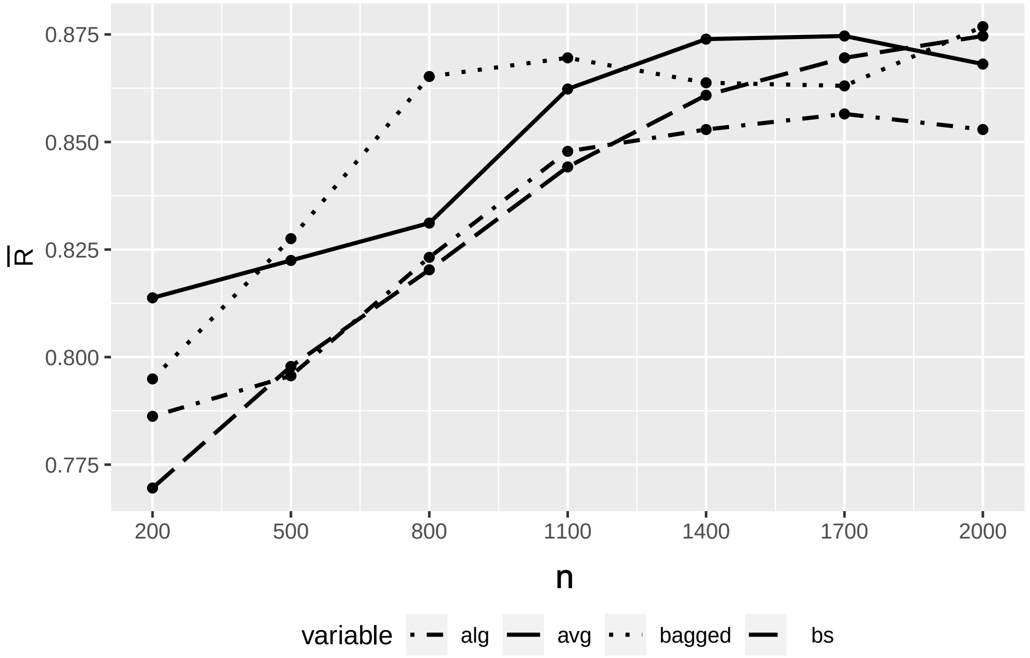

Figure 3.5. Results of the simulations on the *Alarm* network

3.5 Discussion

This work proposed a novel procedure to select an adjustment set for causal effect estimation when the true causal graph is unknown. Firstly bootstrap is used to generate replicates of the original dataset, and then a graph is learnt on each of them. Next, adjustment sets are selected from the obtained graphs according to the back-door criterion and put together, thus forming a multiset. Finally, the set with the greatest multiplicity in the multiset is selected as the procedure's output. This way of using bootstrap is also called bootstrap-aggregating or bagging, and to our knowledge it has never been used before to select an adjustment set directly.

The technique is tested on different networks, and its results are compared to those obtained with three alternative methods. All the chosen procedures show similar levels of accuracy, which generally decrease as the complexity of the considered structure increases. On the simplest graph, at small sample sizes, the bagged adjustment has a lower accuracy than other benchmark methods. However, as the sample size increases, the performance of all the methods, including bagging, improves and seem to be asymptotically equivalent. When considering more complex graphs, the relative performance of the novel procedure stands out. In particular, the results on the most complex diagram show that even if all the methods achieve low accuracy levels, bagging produces the most accurate results at both low and

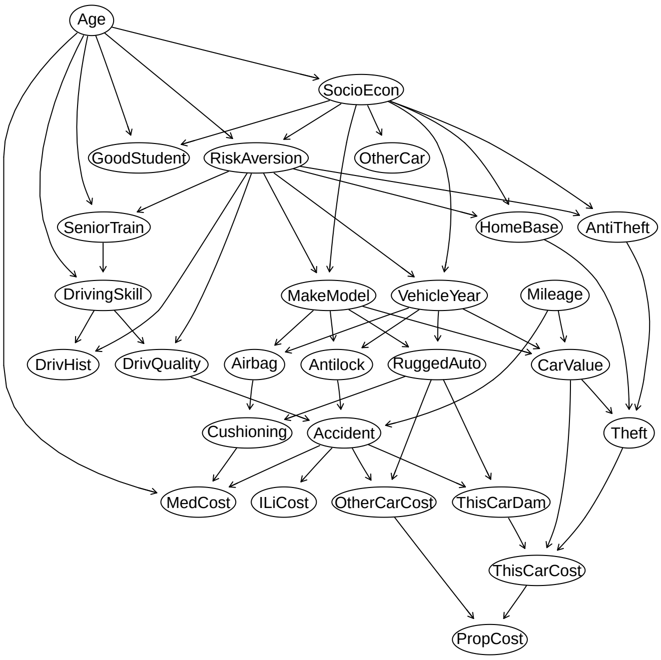

Figure 3.6. The *Insurance* graph

high sample sizes.

When dealing with the estimation of causal effects from observational data and an unknown causal graph, the literature's most common approach consists of learning a causal graph and selecting an adjustment set according to its structure. Much progress has been made in learning algorithms to obtain the structure that encodes the information contained in the data in the best way.

This work contributes to the literature by proposing a procedure that directly aims at selecting a sufficient adjustment set with the use of bootstrap. Bootstrapping has already been used in the graph learning phase to measure confidence toward the presence of an edge between two nodes. However, this implementation is still oriented on recovering the most reliable graph structure and then selecting the adjustment set in a different step. Instead, the proposed procedure uses bootstrap to obtain a multiset of adjustment sets from multiple estimated graphs. This way of proceeding detaches the adjustment set selection from a single learnt graph's structure, thus aiming to achieve higher accuracy.

The effectiveness of the proposed method has been assessed on three discrete graphs of increasing complexity, comparing its results to three benchmark procedures. The findings remain tied to the tested causal graphs and the assumptions made in the simulations. The code of the novel procedure, developed in R Statistical Software, is currently not publicly available. However, future developments of this work could include the creation of an R package to select bootstrap-aggregated

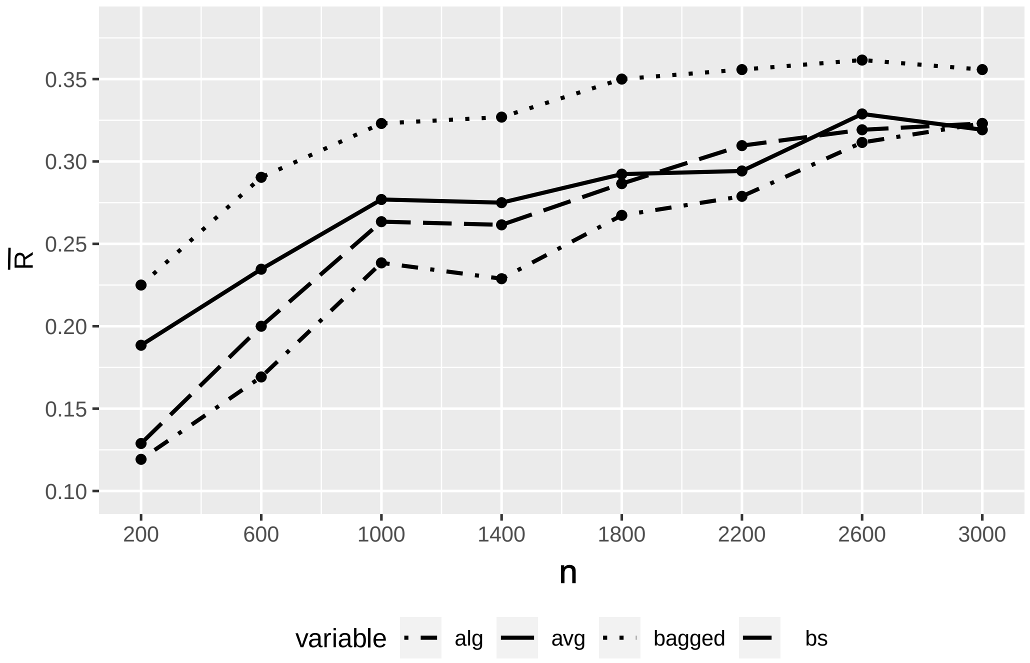

Figure 3.7. Results of the simulations on *Insurance* network

adjustment sets in a fully automated way.

Further analysis is necessary to assess how the procedure behaves with different graphs and assumptions. Nothing prevents the application of the procedure to Gaussian Causal Graphs, and its performance in this context has yet to be evaluated. Moreover, additional analyses are required to study the relationship between the number of bootstrap replicates, performance, and computational time of the novel technique. Even if increasing the number of bootstrap replicates over a certain threshold seems to lengthen computational times considerably with minor accuracy gains, the increase in accuracy could still be relevant, especially when applying the procedure to complex graph structures. Lastly, future research could focus on how the method's accuracy varies with different learning algorithms and adjustment selection criteria. In particular, bootstrap-aggregation for the adjustment set could be implemented with learning algorithms that generate completed partially directed acyclic graphs (CPDAGs) and other classes of graphical models.

Bibliography

[1] Alberto Abadie and Matias D. Cattaneo. “Econometric Methods for Program Evaluation”. In: *Annual Review of Economics* 10.1 (Aug. 2018), pp. 465–503. ISSN: 1941-1383. DOI: 10.1146/annurev-economics-080217-053402.

[2] Alberto Abadie and Jann Spiess. “Robust Post-Matching Inference”. In: *Journal of the American Statistical Association* 0.0 (Oct. 2020), pp. 1–13. ISSN: 0162-1459. DOI: 10.1080/01621459.2020.1840383.

[3] Joshua D. Angrist. “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records”. In: *The American Economic Review* (1990), pp. 313–336.

[4] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics*. Princeton university press, 2008.

[5] Susan Athey and Guido W. Imbens. “The State of Applied Econometrics: Causality and Policy Evaluation”. In: *Journal of Economic Perspectives* 31.2 (May 2017), pp. 3–32. ISSN: 0895-3309. DOI: 10.1257/jep.31.2.3.

[6] Pierluigi Balduzzi et al. *The Economic Effects of COVID-19 and Credit Constraints: Evidence from Italian Firms’ Expectations and Plans*. SSRN Scholarly Paper ID 3682943. Rochester, NY: Social Science Research Network, Aug. 2020.

[7] Elias Bareinboim and Judea Pearl. “Causal Inference and the Data-Fusion Problem”. In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7345–7352.

[8] Jose Maria Barrero, Nicholas Bloom, and Steven J. Davis. *Why Working from Home Will Stick*. Working Paper 28731. National Bureau of Economic Research, Apr. 2021. DOI: 10.3386/w28731.

[9] Alexander W. Bartik et al. *What Jobs Are Being Done at Home During the Covid-19 Crisis? Evidence from Firm-Level Surveys*. Working Paper 27422. National Bureau of Economic Research, June 2020. DOI: 10.3386/w27422.

[10] Ingo A. Beinlich et al. "The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks". In: *AIME 89*. Springer, 1989, pp. 247–256.

[11] Alexander Bick, Adam Blandin, and Karel Mertens. "Work from Home Before and After the COVID-19 Outbreak". In: *Federal Reserve Bank of Dallas, Working Papers* 2020.2017 (Feb. 2021). DOI: 10.24149/wp2017r2.

[12] John Binder et al. "Adaptive Probabilistic Networks with Hidden Variables". In: *Machine Learning* 29.2 (1997), pp. 213–244.

[13] Nicholas Bloom et al. "Does Working from Home Work? Evidence from a Chinese Experiment *". In: *The Quarterly Journal of Economics* 130.1 (Feb. 2015), pp. 165–218. ISSN: 0033-5533. DOI: 10.1093/qje/qju032.

[14] Luca Bonacini, Giovanni Gallo, and Sergio Scicchitano. "Working from Home and Income Inequality: Risks of a 'New Normal' with COVID-19". In: *Journal of Population Economics* 34.1 (Jan. 2021), pp. 303–360. ISSN: 1432-1475. DOI: 10.1007/s00148-020-00800-7.

[15] Simone Borra and Agostino Di Ciaccio. "Improving Nonparametric Regression Methods by Bagging and Boosting". In: *Computational Statistics & Data Analysis*. Nonlinear Methods and Data Mining 38.4 (Feb. 2002), pp. 407–420. ISSN: 0167-9473. DOI: 10.1016/S0167-9473(01)00068-8.

[16] Emanuele Brancati and Raffaele Brancati. *Heterogeneous Shocks in the COVID-19 Pandemic: Panel Evidence from Italian Firms*. SSRN Scholarly Paper ID 3597650. Rochester, NY: Social Science Research Network, May 2020. DOI: 10.2139/ssrn.3597650.

[17] Leo Breiman. "Bagging Predictors". In: *Machine learning* 24.2 (1996), pp. 123–140.

[18] Bradley M Broom, Kim-Anh Do, and Devika Subramanian. "Model Averaging Strategies for Structure Learning in Bayesian Networks with Limited Data". In: *BMC Bioinformatics* 13.S13 (Aug. 2012), S10. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-S13-S10.

[19] Carlos Cinelli, Andrew Forney, and Judea Pearl. "A Crash Course in Good and Bad Controls". In: *SSRN Electronic Journal* (2020). ISSN: 1556-5068. DOI: 10.2139/ssrn.3689437.

[20] Diego Colombo and Marloes H. Maathuis. "Order-Independent Constraint-Based Causal Structure Learning." In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 3741–3782.

[21] Anthony C. Constantinou. “Evaluating Structure Learning Algorithms with a Balanced Scoring Function”. In: *arXiv:1905.12666 [cs, stat]* (Sept. 2020). arXiv: `1905.12666 [cs, stat]`.

[22] David Roxbee Cox. “Planning of Experiments.” In: (1958).

[23] Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, Jan. 2021. ISBN: 978-0-300-25588-1.

[24] Sandrine Dudoit and Jane Fridlyand. “Bagging to Improve the Accuracy of a Clustering Procedure”. In: *Bioinformatics* 19.9 (2003), pp. 1090–1099.

[25] Alan Felstead and Darja Reuschke. *Homeworking in the UK: Before and during the 2020 Lockdown*. https://wiserd.ac.uk/publications/homeworking-uk-and-during-2020-lockdown. Monograph. Aug. 2020.

[26] Ronald A. Fisher. “The Design of Experiments”. In: (1949).

[27] Nir Friedman, Moises Goldszmidt, and Abraham Wyner. “Data Analysis with Bayesian Networks: A Bootstrap Approach”. In: *arXiv preprint arXiv:1301.6695* (2013). arXiv: `1301.6695`.

[28] Maxime Gasse, Alex Aussem, and Haytham Elghazel. “A Hybrid Algorithm for Bayesian Network Structure Learning with Application to Multi-Label Learning”. In: *Expert Systems with Applications* 41.15 (2014), pp. 6755–6772.

[29] Fred Glover. “Future Paths for Integer Programming and Links to Artificial Intelligence”. In: *Computers & operations research* 13.5 (1986), pp. 533–549.

[30] Clark Glymour, Peter Spirtes, and Richard Scheines. “Causal Inference”. In: *Erkenntnis* 35.1-3 (1991), pp. 151–189.

[31] Sander Greenland, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research”. In: *Epidemiology* (1999), pp. 37–48.

[32] Kyoungnam Ha, Sungzoon Cho, and Douglas MacLachlan. “Response Models Based on Bagging Neural Networks”. In: *Journal of Interactive Marketing* 19.1 (2005), pp. 17–30.

[33] Trygve Haavelmo. “The Statistical Implications of a System of Simultaneous Equations”. In: *Econometrica, Journal of the Econometric Society* (1943), pp. 1–12.

[34] Ben B. Hansen. “Full Matching in an Observational Study of Coaching for the SAT”. In: *Journal of the American Statistical Association* 99.467 (2004), pp. 609–618.

[35] Paul W. Holland. “Statistics and Causal Inference”. In: *Journal of the American statistical Association* 81.396 (1986), pp. 945–960.

[36] Yimin Huang and Marco Valtorta. "Pearl's Calculus of Intervention Is Complete". In: *arXiv preprint arXiv:1206.6831* (2012). arXiv: `1206.6831`.

[37] Paul Hünermund and Elias Bareinboim. "Causal Inference and Data Fusion in Econometrics". In: *arXiv preprint arXiv:1912.09104* (2019). arXiv: `1912.09104`.

[38] Guido W. Imbens. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review". In: *Review of Economics and statistics* 86.1 (2004), pp. 4–29.

[39] Guido W. Imbens. "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics". In: *Journal of Economic Literature* 58.4 (2020), pp. 1129–1179.

[40] Guido W. Imbens and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences.* Cambridge University Press, 2015.

[41] Seiya Imoto et al. "Bootstrap Analysis of Gene Networks Based on Bayesian Networks and Nonparametric Regression". In: *Genome Informatics* 13 (2002), pp. 369–370.

[42] Uffe B. Kjaerulff and Anders L. Madsen. "Bayesian Networks and Influence Diagrams". In: *Springer Science+ Business Media* 200 (2008), p. 114.

[43] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT press, 2009.

[44] Steffen L. Lauritzen and David J. Spiegelhalter. "Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 50.2 (1988), pp. 157–194.

[45] Daniela Marella and Paola Vicard. "Bayesian Network Structural Learning from Complex Survey Data: A Resampling Based Approach". In: *Statistical Methods & Applications* (Jan. 2022). ISSN: 1613-981X. DOI: `10.1007/s10260-021-00618-x`.

[46] Dimitris Margaritis. *Learning Bayesian Network Model Structure from Data.* Tech. rep. Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2003.

[47] Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference.* Cambridge University Press, 2015.

[48] Jerzy Neyman and Karolina Iwaszkiewicz. "Statistical Problems in Agricultural Experimentation". In: *Supplement to the Journal of the Royal Statistical Society* 2.2 (1935), pp. 107–180.

[49] Judea Pearl. “Causal Diagrams for Empirical Research”. In: *Biometrika* 82.4 (Dec. 1995), pp. 669–688.

[50] Judea Pearl. “Causal Inference in Statistics: An Overview”. In: *Statistics Surveys* 3.none (Jan. 2009). ISSN: 1935-7516. DOI: `10.1214/09-SS057`.

[51] Judea Pearl. *Models, Reasoning and Inference*. Vol. 19. 2000.

[52] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Rev. 2. ed., transferred to digital printing. The Morgan Kaufmann Series in Representation and Reasoning. San Francisco, Calif: Morgan Kaufmann, 2009. ISBN: 978-1-55860-479-7.

[53] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. Chichester, West Sussex, United Kingdom: Wiley, 2016. ISBN: 978-1-119-18684-7 978-1-119-18685-4.

[54] Emilija Perković et al. “A Complete Generalized Adjustment Criterion”. In: *Uncertainty in Artificial Intelligence* (2015).

[55] R. Core Team. “R: A Language and Environment for Statistical Computing”. In: (2013).

[56] Paul Rosenbaum. *Observation and Experiment*. Harvard University Press, 2018.

[57] Paul R. Rosenbaum. “A Characterization of Optimal Designs for Observational Studies”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3 (1991), pp. 597–610.

[58] Donald B. Rubin. “Causal Inference Using Potential Outcomes: Design, Modeling, Decisions”. In: *Journal of the American Statistical Association* 100.469 (2005), pp. 322–331.

[59] Donald B. Rubin. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” In: *Journal of educational Psychology* 66.5 (1974), p. 688.

[60] S Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach. Prentice Hall*. Jan. 2009.

[61] Stuart Russell and Peter Norvig. “Artificial Intelligence: A Modern Approach”. In: (2002).

[62] Marco Scutari. “Learning Bayesian Networks with the Bnlearn R Package”. In: *Journal of Statistical Software* 35 (July 2010), pp. 1–22. ISSN: 1548-7660. DOI: `10.18637/jss.v035.i03`.

[63] Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. "Who Learns Better Bayesian Network Structures: Accuracy and Speed of Structure Learning Algorithms". In: *arXiv:1805.11908 [stat]* (July 2019). arXiv: `1805.11908 [stat]`.

[64] Jonathan M. Snowden, Sherri Rose, and Kathleen M. Mortimer. "Implementation of G-computation on a Simulated Data Set: Demonstration of a Causal Inference Technique". In: *American journal of epidemiology* 173.7 (2011), pp. 731–738.

[65] Peter Spirtes et al. *Causation, Prediction, and Search.* MIT press, 2000.

[66] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." In: *Statistical Science* 5.4 (1990), pp. 465–472.

[67] Elizabeth A. Stuart and Kerry M. Green. "Using Full Matching to Estimate Causal Effects in Nonexperimental Studies: Examining the Relationship between Adolescent Marijuana Use and Adult Outcomes." In: *Developmental psychology* 44.2 (2008), p. 395.

[68] Elizabeth A. Stuart, Gary King, et al. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference". In: *Journal of statistical software* (2011).

[69] Johannes Textor et al. "Robust Causal Inference Using Directed Acyclic Graphs: The R Package 'Dagitty'". In: *International journal of epidemiology* 45.6 (2016), pp. 1887–1894.

[70] Jin Tian and Judea Pearl. "A General Identification Condition for Causal Effects". In: *Aaai/Iaai.* 2002, pp. 567–573.

[71] Robert J. Tibshirani and Bradley Efron. "An Introduction to the Bootstrap". In: *Monographs on statistics and applied probability* 57 (1993), pp. 1–436.

[72] Jan Tinbergen. "Determination and Interpretation of Supply Curves: An Example". In: *Zeitschrift fur Nationalokonomie* 1.5 (1930), pp. 669–679.

[73] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. "The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm". In: *Machine learning* 65.1 (2006), pp. 31–78.

[74] Thomas Verma and Judea Pearl. "Causal Networks: Semantics and Expressiveness". In: *Machine Intelligence and Pattern Recognition.* Vol. 9. Elsevier, 1990, pp. 69–76.

[75] Janine Witte et al. "On Efficient Adjustment in Causal Graphs". In: *Journal of Machine Learning Research* 21 (2020), p. 246.

[76] Sewall Wright. “Correlation and Causation”. In: (1921), pp. 557–585.

[77] Sewall Wright. “Systems of Mating. I. The Biometric Relations between Parent and Offspring”. In: *Genetics* 6.2 (1921), p. 111.